PRENTICE-HALL EE SERIES

Le Croissette

Transistors

621.38151
L469t
c.2

University of Florida Libraries

Engineering & Physics Library

The Gift of

Eugene Chenette

Prentice-Hall, Inc.

Digitized by the Internet Archive
in 2013

http://archive.org/details/transistors00lecr

Transistors

by

Dennis Le Croissette

Jet Propulsion Laboratory
California Institute of Technology

Prentice-Hall, Inc.

Englewood Cliffs, N. J.

1963

PRENTICE-HALL INTERNATIONAL, INC., *London*
PRENTICE-HALL OF AUSTRALIA, PTY., LTD., *Sydney*
PRENTICE-HALL OF FRANCE, S.A.R.L., *Paris*
PRENTICE-HALL OF JAPAN, INC., *Tokyo*
PRENTICE-HALL DE MEXICO, S.A., *Mexico City*
PRENTICE-HALL OF CANADA, LTD., *Toronto*

Prentice-Hall Electrical Engineering Series

William L. Everitt, *Editor*

© 1963 by Prentice-Hall, Inc., Englewood Cliffs, N.J.
All rights reserved. No part of this book may be reproduced in any form, by mimeograph or any other means, without permission in writing from the publisher.

Library of Congress Catalog Card Number: 63-7561

Printed in the United States of America 93017-C

Preface

The aim of this book is to present and explain the basic principles of semiconductor electronics as they apply to junction transistors. The introductory approach makes it suitable for use as an undergraduate text in electrical engineering. It will also be helpful to engineers in industry who require a knowledge of transistors.

The fundamentals of semiconductor devices are introduced here from a physical viewpoint. In my opinion, this is the best preparation for later work in electronic circuitry where device limitations must be considered. Moreover, vacuum tube

electronics is relatively easy to understand after these concepts have been developed. This opinion is supported by experience at Drexel Institute of Technology where part of this material was used as the first course in electronics.

Part 1 of this book is concerned with the flow of electrons and holes in semiconducting materials, leading to an understanding of current flow in *p-n* junctions and transistors.

In Part 2, the "black box" concept of the transistor is developed. Analysis is restricted to low frequency operation of the device, and hybrid (or *h*) parameter representation is used. Direct-current operation of the transistor is also considered.

Part 3 consists mainly of an analysis of the high frequency and pulse operation of the transistor. The hybrid-π circuit is derived and used to determine the high frequency response of the common emitter stage. Pulse operation is described from the charge storage point of view. In conclusion, the tunnel diode is discussed.

It is apparent that this book, designed primarily for engineers, cannot derive all the physical relationships necessary for an analysis of the transistor. The references at the end of each chapter assist the reader to find the derivations he requires.

I feel strongly that this textbook is not suitable for a course oriented towards solving numerical problems. Therefore, many descriptive questions have been given to encourage students to review and rephrase the material.

I would like to acknowledge the assistance of John Yarwood of London, England and Eugene J. Rosenbaum of Philadelphia, Pa. who read part of the manuscript while it was in preparation. My former colleagues at Drexel Institute of Technology have made a number of helpful comments. I am also grateful to Albert F. Fuchs for contributing many of the problems and to Mrs. Noreen J. Reyes for typing most of the manuscript.

DENNIS LE CROISSETTE

Contents

PART ONE: THE MOTION OF ELECTRONS AND HOLES IN SEMICONDUCTING MATERIALS

1 **Electrical Conduction in Semiconductors** 3

1-1 Introduction, 4. 1-2 Simple theory of the hydrogen atom, 4. 1-3 Bohr's theory of the hydrogen atom, 5. 1-4 Energy of the electron orbits of the hydrogen atom, 7. 1-5 Transitions between the orbits, 9. 1-6 Energy level representation, 10. 1-7 Limitations of the Bohr theory, 11. 1-8 The wave nature of matter, 11. 1-9 Wave-mechanical theory of atomic structure, 14. 1-10 The atomic table, 15. 1-11 Chemical valence and atomic binding, 17. 1-12 Crystals, 18. 1-13 Electron energy levels in crystals: band theory, 20. 1-14 Insulators, 24. 1-15 Conductors, 25. 1-16 Semiconductors, 25. 1-17 Electrons and positive holes, 26. 1-18 Mobility, 28. 1-19 *p*- and *n*-type semiconductors, 29. 1-20 Generation and recombination, 35.

2 **Electrons and Holes in Semiconductors** 43

2-1 The density of states in an energy band, 44. 2-2 Energy state densities in semiconductors and the effect of temperature, 47. 2-3 Intrinsic semiconductor, 51. 2-4 *n*-type semiconductor, 54. 2-5 *p*-type semiconductor, 55. 2-6 Variation of the Fermi level with temperature, 56. 2-7 The law of mass action, 57. 2-8 Mobile and immobile charges, 58. 2-9 The Hall effect, 61.

3 **Junctions between Materials** 67

3-1 Work function in a metal, 68. 3-2 Junctions between metals, 70. 3-3 Semiconductor junctions with no applied voltage, 72. 3-4 Semiconductor junction with applied voltage: junction capacitance, 79. 3-5 Flow of current across a *p-n* junction: the rectifier equation, 82. 3-6 The breakdown region of a *p-n* junction, 86.

4 **The Continuity Equation** 89

4-1 Derivation of the continuity equation, 90. 4-2 Application of the continuity equations to the abrupt *p-n* junction operating with a constant current, 94. 4-3 Calculation of the charge densities at the edge of the depletion regions, 99. 4-4 Solution of the continuity equations for a reverse-biased abrupt *p-n* junction, 101. 4-5 Solution of the continuity equations for a forward-biased abrupt *p-n* junction, 104. 4-6 A brief discussion of the *p-n* junction, 107. 4-7 The incremental resistance of a forward-biased *p-n* junction, 108. 4-8 Einstein equation, 108.

5 **The Junction Transistor** 112

5-1 The *npn* transistor, 113. 5-2 The *pnp* transistor, 116. 5-3 Characteristics of the current flow across the base region, 117. 5-4 The transistor as a device: common base circuit, 123. 5-5 The transistor as a device: common emitter circuit, 126. 5-6 The manufacture of transistors, 128.

PART TWO: TRANSISTORS AT LOW FREQUENCIES AND UNDER D-C CONDITIONS

6 **The Transistor as a Two Port Network** 135

6-1 General considerations of electronic devices as amplifiers, 136. 6-2 Alternating-current operation of the transistor, 139. 6-3 *z*, *y*, and *h* parameters: general comments, 142. 6-4 Equivalent circuit using *z* parameters: the T circuit, 144. 6-5 Equivalent circuit using *y* parameters: the π circuit, 148. 6-6 Equivalent circuit using *h* parameters: a two generator representation, 149. 6-7 Low frequency measurements on transistors, 150.

7 **Low Frequency *h* Parameter Representation** 154

7-1 The small-signal *h* parameters, 155. 7-2 Analysis of the general *h* parameter equivalent circuit, 156. 7-3 Relationship of the three sets of *h* parameters, 159. 7-4 Comparison of the three configurations, 161. 7-5 Variation in *h* parameters with I_E and V_{CE}, 163.

8 **Single and Multistage a-c Amplifiers** 167

8-1 The characteristic curves, 168. 8-2 The load line, 170. 8-3 Operating point stability, 172. 8-4 Practical biasing circuits, 178. 8-5 Power amplifiers, 182. 8-6 Power amplifiers: practical limitations, 187. 8-7 Thermal runaway, 189. 8-8 Multistage amplifier, 191. 8-9 Frequency response of the multistage amplifier, 194.

PART THREE: FURTHER THEORY OF THE TRANSISTOR, AND THE TUNNEL DIODE

9 **Physical Characteristics of the Transistor** 199

9-1 Solution of the continuity equation in the base region for direct currents, 200. 9-2 Components of alpha, 202. 9-3 The intrinsic transistor, 203. 9-4 The continuity equation in the base region for small alternating signals, 204. 9-5 Base width modulation: transmission line analogy, 208. 9-6 The hybrid-π representation, 211. 9-7 Variation of alpha with emitter current, 217. 9-8 The drift transistor, 220. 9-9 Transistor noise, 223. 9-10 Surface effects in transistors, 225. 9-11 Punch-through, 226. 9-12 Epitaxial films, 226.

10 **High Frequency and Pulse Operation of the Transistor** 229

10-1 The hybrid-π circuit, 230. 10-2 The Miller effect, 231. 10-3 Frequency response of the transistor in the common emitter connection: hybrid-π model, 233. 10-4 Frequency response of the iterative common emitter stage: hybrid-π model, 235. 10-5 Frequency response of the transistor in the common base connection: simple model, 236. 10-6 Comparison of the hybrid-π circuit with the actual transistor, 238. 10-7 Saturation, 242. 10-8 Switching the transistor on: use of the hybrid-π model, 244. 10-9 Definition of switching times for the transistor: saturation case, 247. 10-10 Charge control of the transistor, 249. 10-11 Switching times for a saturated transistor, 254.

11 **The Tunnel Diode** 260

11-1 Quantum mechanical tunneling, 261. 11-2 Energy bands of the tunnel diode, 263. 11-3 Forward and reverse characteristics, 265. 11-4 The tunnel diode as a circuit element, 268.

APPENDIX 274

INDEX 275

PART 1

The Motion of Electrons and Holes in Semiconducting Materials

1

Electrical Conduction in Semiconductors

When the vacuum tube was the only practicable electronic amplifier, the subject of electronics was predominantly concerned with the flow of electrons in a vacuum or in a gas at low pressure. In the past twenty years, however, considerable attention has been focused on the conduction of electricity in solid materials. Research in solid-state electronics was accelerated by the development of the transistor in 1948. Since that date, a study of the processes governing current flow in semiconductors has been essential for an understanding of electronics.

This chapter begins with a review of the basic atomic properties of matter leading to the energy band representation of insulators, conductors, and semiconductors. Later in this chapter, the fundamental concept of hole and electron conduction in a semiconductor is developed, and recombination between holes and electrons is briefly discussed.

1–1 Introduction

In the past, solid materials were divided into two main groups according to their electrical resistivity: *conductors* and *insulators*. Conductors were loosely defined as having resistivities less than about 10^{-3} ohm cm, and this group of materials included all the metals. Insulators were considered to include substances with resistivities higher than 10^{8} ohm cm and in this category were ebonite, glass, mica, etc. Materials having a resistivity between these two values were known, but it was often difficult to measure their resistivity accurately since different specimens of the same substance gave different results. Such materials are called *semiconductors*, and an explanation of their electrical properties will be given later in this chapter. The elements *germanium* and *silicon* are the two most used semiconductors, and the electrical properties of these materials are responsible for the operation of transistors and other semiconducting devices.

The transistor is the most important of the *solid-state devices*. In a solid, the passage of current is determined by the binding forces which act between electrons and the positively charged atomic nuclei. Therefore, the emphasis in beginning this study of modern electronic theory will be placed upon an understanding of the physical properties of the atom.

1–2 Simple Theory of the Hydrogen Atom

Rutherford first proposed an atomic model of the type shown in Fig. 1–1. In this diagram, an electron is shown revolving in a circular orbit around a positively charged nucleus. In the case of hydrogen, it is known that the positive charge in the nucleus is numerically equal to the charge on the

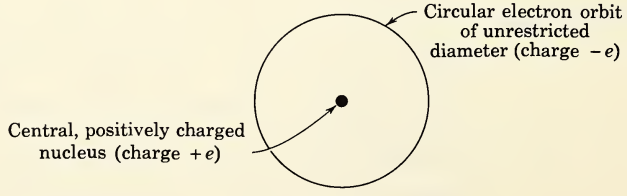


FIG. 1–1. Simple picture of the hydrogen atom.

electron (e). Coulomb's law states that the force of attraction between these two charges is

$$\frac{e^2}{4\pi\varepsilon_0 r^2}$$

where r is the radius of the orbit in meters, and ε_0 is the permittivity of a vacuum.

If the electron is revolving with velocity v in meters per second, the radial force necessary to restrain it in a circular orbit of radius r is mv^2/r. Hence,

$$\frac{e^2}{4\pi\varepsilon_0 r^2} = \frac{mv^2}{r} \tag{1-1}$$

Equation (1-1) is seen to be a result of the application of the classical laws of mechanics and electrostatics. It does not place any restriction on the radius of the electron orbit, and it is inadequate to explain why the electron remains in equilibrium in its path around the nucleus. Any change in the radius of the orbit will be accompanied by a change in the energy of the system [see Eq. (1-10)]. According to this theory, if the electron moves to a new orbit closer to the central nucleus, the energy of the system falls, and the excess energy is radiated out into space. This process will be repeated as time goes on and eventually the electron will fall into the nucleus. It is clear that Eq. (1-1) alone is incapable of explaining the motion of the electron around the nucleus.

An additional discrepancy between this simple theory and experiment is that this model does not explain the mechanism of the emission and absorption of radiation which takes place in a gas. Studies of the emission and absorption spectra of various gases show that certain frequencies appear which are characteristic of the gas. It is concluded, therefore, that definite energy levels are possible in the atom and that emission or absorption of radiation occurs when electrons move from one energy level to another. Classical theory is unable to explain this phenomenon since it does not indicate the presence of definite electron energy levels in the atom.

1-3 Bohr's Theory of the Hydrogen Atom

In 1913, Bohr suggested that these difficulties could be overcome by a direct application of Planck's *quantum theory*. The idea of a *quantum of energy* is fundamental to modern electronic theory since it is concerned with the emission and absorption of energy on an atomic scale. This was one of the first of many successful applications of quantum theory in the solution of atomic problems and led to the later and more exact theory of the atom.

The quantum theory states that all energy is emitted and absorbed in the form of multiples of a unit called the *quantum*. Unlike the more familiar

units of energy, the energy in each quantum is a function of frequency and is given by

$$E = h\nu \tag{1–2}$$

where E is the energy of one quantum, in joules; h is Planck's constant, 6.62×10^{-34} joule sec; and ν is the frequency of radiation, in cycles per second. The amount of energy in a quantum is very small so that the emission and absorption of energy in experiments in ordinary mechanics are not affected. However, many atomic phenomena are governed by laws which may be deduced from the quantum theory.

Bohr applied this theory to show that it was possible to restrict the orbital radii to certain discrete values. He imposed a quantum condition on the value of the angular momentum in an orbit and showed that this was necessary for the establishment of equations which fitted the experimental facts. When the magnitude of any quantity is restricted to multiples of a small unit (such as a quantum of energy), the quantity is said to be "quantized." The quantization carried out by Bohr [Eq. (1–5)] was later shown to be the restriction of angular momentum to multiples of a small unit.

Bohr's model of the atom is based upon two assumptions.

1. The radius of the orbit is limited to certain fixed values, and no emission or absorption of energy takes place while the electron remains in one of the permitted orbits.
2. When the electron jumps from an orbit of one allowed radius to another, emission or absorption of a quantum of energy occurs. The frequency of radiation can be found from

$$h\nu = E_2 - E_1 \tag{1–3}$$

where E_2 and E_1 are the energy levels of the two electron orbits.

Assuming that the electron revolves in a circle around the nucleus, its angular momentum is

$$mvr \tag{1–4}$$

The quantum condition imposed by Bohr on the angular momentum was that

$$mvr = \frac{nh}{2\pi} \tag{1–5}$$

where n is a positive integer known as the principal quantum number. This number determines the energy of the orbit.

Eliminating v from Eqs. (1–1) and (1–5) gives

$$r_n = \frac{n^2h^2\varepsilon_0}{\pi me^2} \tag{1–6}$$

where r_n is the radius of the nth orbit.

Inserting the known values of the physical constants, h, ε_0, m, and e (see Appendix) in Eq. (1–6) and taking $n = 1$

$$r_1 = 0.529 \times 10^{-10} \text{ meter}$$
$$= 0.529 \text{ A} \quad \text{(Angstrom units)} \tag{1–7}$$

The value of r_1 for the hydrogen atom is known as the Bohr radius, r_B. It is of the same order as the effective radius of the atom found from kinetic theory. From Eq. (1–6),

$$r_1 = r_B$$
$$r_2 = 4r_B$$
$$r_3 = 9r_B$$
$$r_4 = 16r_B \quad \text{etc.} \tag{1–8}$$

In this way, Bohr suggested that the allowed radii that the electron orbits could take were in the ratio $1^2, 2^2, 3^2, 4^2, \ldots$. This is shown in Fig. 1–2.

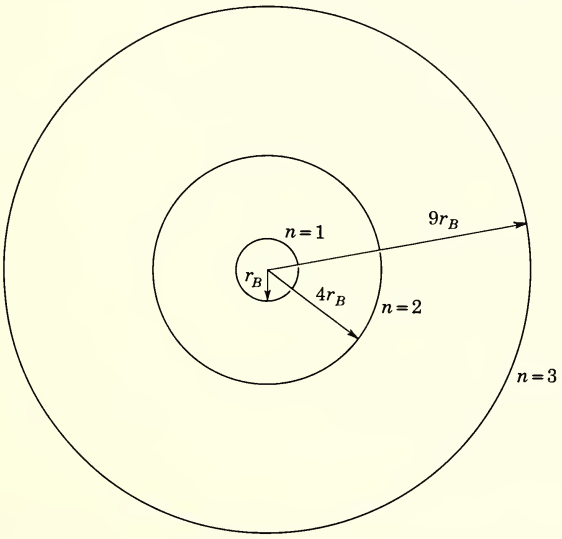


FIG. 1–2. Bohr theory of the hydrogen atom showing the first three electron orbits.

1–4 Energy of the Electron Orbits of the Hydrogen Atom

The kinetic energy of an electron revolving in an orbit with velocity v is $\frac{1}{2}mv^2$. From Eq. (1–1),

$$\text{K.E.} = \frac{e^2}{8\pi\varepsilon_0 r_n} \tag{1–9}$$

The potential energy of this electron is the product of the electric potential at the point and the charge on the electron.

$$\text{P.E.} = \frac{e}{4\pi\varepsilon_0 r_n}(-e) = \frac{-e^2}{4\pi\varepsilon_0 r_n} \tag{1-10}$$

Note that the potential energy is negative and numerically is twice the value of the kinetic energy. In electrical systems, the potential energy at a point may be negative without any special significance being attached to the fact. Electrical potential is defined as the work done against the field in bringing a unit positive charge from infinity up to the point; the electrical

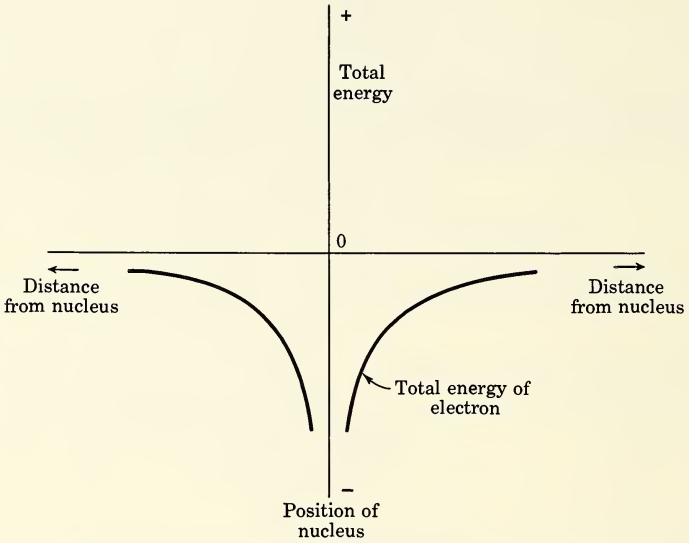


FIG. 1–3. Plot of the total energy of an electron close to the hydrogen nucleus.

potential at a point near a negative charge is, therefore, negative in sign. This definition means that the potential energy of the ionized atom is taken to be zero (see Fig. 1–3).

The total energy of the electron in orbit of radius r_n is

$$\text{K.E.} + \text{P.E.} = -\frac{e^2}{8\pi\varepsilon_0 r_n} \tag{1-11}$$

Substituting for r_n from Eq. (1–6),

$$\text{Total energy, } E_n = -\frac{me^4}{8n^2h^2\varepsilon_0^2} \tag{1-12}$$

where E_n is the energy of the nth orbital state (radius r_n).

The inner orbit where $n = 1$ has the lowest (i.e., the most negative) energy and is known as the *ground state*. In the hydrogen atom, the electron will remain in this state unless it is disturbed by outside influences.

1-5 Transitions between the Orbits

The simple case of hydrogen serves as a useful starting point towards understanding the way in which the absorption and emission of energy take place

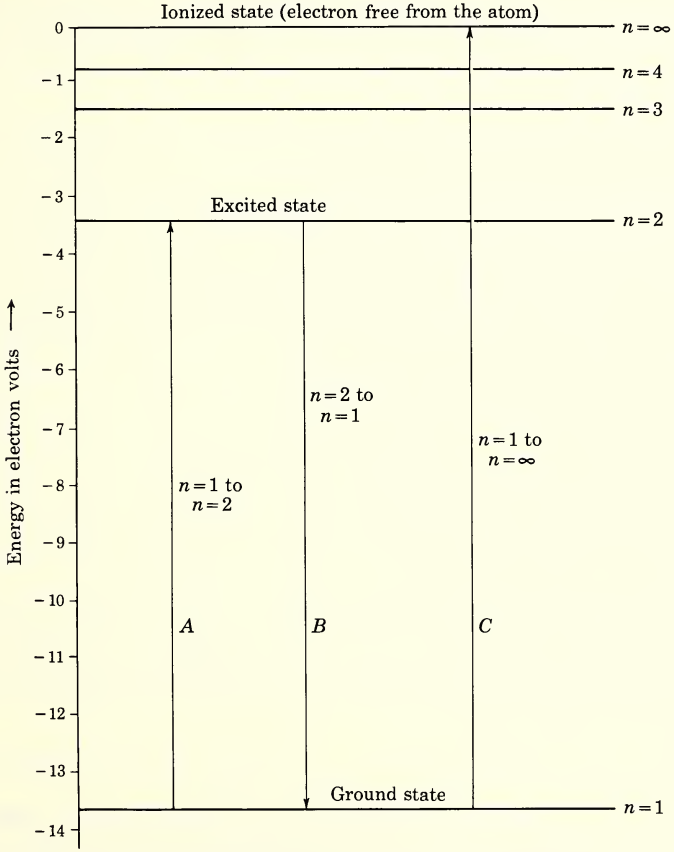


FIG. 1-4. Energy level representation for the Bohr theory of the hydrogen atom.

in an atomic system. The single orbital electron of a hydrogen atom can be removed from the ground state $n = 1$ and brought into the state $n = 2$ if electromagnetic radiation of a certain frequency is incident on the atom. To move the electron from one orbit to another requires a quantum of energy

equal to the energy difference of the two orbits. A quantum of electromagnetic energy such as is absorbed by one atom in this process is referred to as a *photon of radiation* and the frequency of the photon may be calculated as before from

$$h\nu = E_2 - E_1 \tag{1–3}$$

It is found that a photon of energy of this wavelength is absorbed by only one atom causing the electron to move from $n = 1$ to $n = 2$. In this sense, the photon of radiation acts as if it were concentrated at a point in apparent contradiction to the usual concept of wave motion. This particle-like nature of a photon is a concept which has been extended and developed to provide the basis for more advanced work on quantum mechanics which will be briefly discussed later.

The atom is said to be *excited* when the electron is raised to a new orbit. If the electron is given sufficient energy to remove it completely from the influence of the nucleus, the atom is *ionized*, and the energy necessary to release an electron in this way is called the *ionization energy*, E_I. The ionization energy may be calculated in the case of hydrogen by considering the difference in the energy between orbits $n = 1$ and $n = \infty$. The quantum number ∞ means that the electron is completely removed from the nucleus From Eq. (1–12),

$$\begin{aligned} E_I &= E_\infty - E_1 \\ &= -\frac{me^4}{8h^2\varepsilon_0^2}\left[\frac{1}{\infty} - \frac{1}{1}\right] = \frac{me^4}{8h^2\varepsilon_0^2} \end{aligned} \tag{1–13}$$

Putting in the known values of the physical constants,

$$E = 13.60 \text{ ev} \quad \text{(electron volts)} \tag{1–14}$$

where one *electron volt* is the energy which an electron acquires in moving through a potential difference of one volt and equals 1.6×10^{-19} joule. Equation (1–14) is in agreement with the experimentally measured values.

1–6 Energy Level Representation

Figure 1–4 shows a representation of the energy levels of the first four orbits in the hydrogen atom as predicted by the Bohr theory. The energy of the electron states is plotted vertically with the positively increasing direction upwards and with zero energy at the top. The horizontal axis has no significance here, but in diagrams of this nature which will be used later, it will indicate distance in a material if the energy of the states varies for any reason through the substance.

The ground state for which $n = 1$ is normally occupied. If radiation of frequency given by Eq. (1–3) is incident on the atom, the electron will move into orbit $n = 2$. This transition is indicated by the line A. After remaining

in the excited state for about 10^{-8} second, the reverse transition B will occur, and energy in the form of a photon of frequency ν will be emitted from the atom. The transition C indicates ionization of the atom.

This method of indicating the energy levels of the various possible electron orbits is a most important one and will be used later not only for gases but also for solid-state materials.

1-7 Limitations of the Bohr Theory

The Bohr theory was the first step towards understanding the structure of the atom as it was able to explain the presence of certain frequencies in the emission and absorption spectrum of hydrogen which spectroscopists had already observed. The theory also predicted that other frequencies would be found in the spectrum corresponding to the transitions between various orbits as indicated by Eq. (1–3); these spectral lines were later detected. However, Bohr's theory of the atom was found to be limited to the hydrogen atom and to other atoms which contained only one electron, such as singly ionized helium and doubly ionized lithium. All attempts to apply the theory to many-electron atoms failed to predict the atomic energy levels which were known to exist.

Between 1913 and 1925, many attempts were made by Bohr, Sommerfeld, and others to revise the simple theory so that it more accurately predicted the results of experiments. The principal modification was to assume that the electrons were traveling in elliptical orbits; extra quantum numbers were assigned to limit the value of the eccentricity of each orbit. These attempts, although partially successful, were completely superseded in 1925 by the wave-mechanical approach.

In spite of these difficulties, the Bohr theory was an outstanding step forward in understanding energy levels, and the ideas which have been presented here are fundamental to an understanding of the mechanism of electrical conduction and electronic phenomena in both gases and solids.

1-8 The Wave Nature of Matter

The problem of a more accurate calculation of the energy levels in atoms was effectively solved between 1925 and 1930 as a result of a growing awareness of the duality between radiation and matter. The "particle-like" nature of a photon of radiation has been mentioned previously in connection with the absorption of a photon by an atom. The relatively successful theory of the atom due to Bohr, and other theories of the photoelectric effect, can only be explained by considering the photon to be localized over a small region. Thus the photon was considered to be of the nature of a particle, even though it was difficult to reconcile the conflicting properties of wave motion and those of a particle.

In 1922, de Broglie suggested that the converse of this effect also applies, that is, that a small particle (such as an electron) can be represented by a wave motion. He calculated that the wavelength, λ, associated with a particle of mass m traveling at a velocity v is given by h/mv. Davisson and Germer, and also G. P. Thomson, performed experiments in 1927 to verify this and showed that there were diffraction phenomena associated with the passage of an electron beam into matter. Following their work, the wavelike properties of matter were generally accepted. The wavelength of the radiation varies with the velocity (or energy) of the beam. As an example, substitution of the

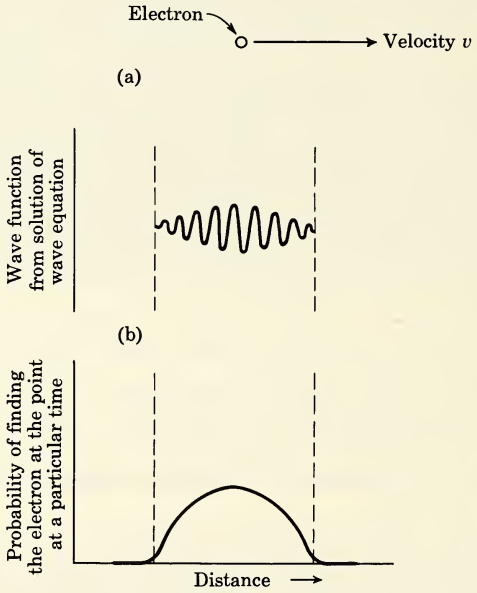


FIG. 1–5. Simplified picture of the solution of the wave equation for a moving electron.

known values of the physical constants shows that $\lambda = 1.22$ A for electrons of energy 100 ev.

Between 1926 and 1930, Schrödinger, Dirac, Heisenberg, and others developed the theoretical implications of the wave properties of matter and produced the mathematical analysis known as *quantum* (or wave) *mechanics*. As one of the first steps in this theory, Schrödinger developed the wave equation which now bears his name and which is an important tool in the analysis of atomic systems. In its simplest form, Schrödinger's wave equation may be written as

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{8\pi^2 m}{h^2}(E - U)\psi = 0 \qquad (1\text{–}15)$$

where E is the total energy, and U is the potential energy as a function of

distance. The solution of this equation gives the *wave function*, ψ, in the x direction. The normalized value of ψ^2 represents the probability of finding the particle at a given point.

A simple example of the solution of Schrödinger's wave equation for the case of an electron moving freely in space is given in Fig. 1–5. The classical picture of a particle moving from left to right with velocity v is shown in Fig. 1–5(a). Part (b) of the figure shows the solution of the wave equation and gives a plot of the wave function, ψ, versus distance. The electron is represented by a group of waves as opposed to a continuous wave train, and this group of waves is often known as a "wave packet." Comparison of parts (a) and (b) of Fig. 1–5 shows that the solution of the wave equation spreads the effect of the electron over a region whereas the classical model limits the electron to a point. It is found that the best description of the position of the electron can be given by specifying the probability of finding the electron at any point. This probability is plotted as a function of distance in Fig. 1–5(c). Although this probability reaches a maximum, it is not possible to locate the electron precisely since there is some probability of finding the electron at any point in the region covered by the wave function of part (b).

When only one electron is concerned in a problem, the concept of the probability of finding it at a given position is a difficult one. However, the concept becomes more understandable when many particles are present since the probability factor is then the fraction of the total number of particles involved. In this case, the probability (which may be calculated directly from quantum mechanics) gives statistical information. In the analysis of atomic systems, such statistical information is of great importance in determining the properties of the system since, even though the probability for a process may be small, the total number of particles may be so high that the number involved in the process is considerable. In addition, it was noted that the quantum mechanical solution for a particle in free space spreads the effect of the particle over a wider region than can be predicted by using classical mechanics (Fig. 1–5). It is also found that certain processes, although classically impossible, become statistically possible using quantum mechanics. In other words, using quantum mechanics, a finite (but perhaps small) probability may be predicted for some phenomena or processes that are completely impossible according to the classical laws of physics. Electrical conduction in semiconductors is based upon such processes.

One of the most interesting and fundamental ideas that has been developed from this approach is the relationship known as Heisenberg's *uncertainty principle*. It is found that there is a limit to the accuracy with which we can determine the position and the momentum of a particle by measuring these quantities simultaneously. If x represents the position and p the momentum of a particle, and Δx and Δp represent the accuracy with which these quantities can be specified,

$$\Delta p \, \Delta x \sim h \tag{1–16}$$

This relationship introduces the concept of *uncertainty* or *indeterminancy* into atomic physics. It means that for experiments which are conducted on an atomic scale, it is not possible to assign a definite position and momentum to a particle at the same time. The more accurately one of these quantities is determined, the less accurately the other one is known. Because of the small size of Planck's constant, h, this effect is important only in atomic phenomena and not in large-scale experiments in the laboratory. In this book, Heisenberg's uncertainty principle is used to determine the properties of the conduction electrons of a metal in Sec. 2–1. It is of great significance in any experiment or analysis where the energy and position of small particles are concerned.

1–9 Wave-mechanical Theory of Atomic Structure

One of the first applications of quantum mechanics was to provide a solution to the problem of the structure of the atom. It was found that the electron energy states require four quantum numbers to describe them.

1. The *principal quantum number*, n, is an integer 1, 2, 3, 4, . . . which mainly determines the energy of the electron state. It corresponds to the quantum number n in the Bohr theory, but the radius of the electron orbit can no longer be stated with accuracy since the probability of finding the electron is finite over a considerable region around the nucleus. The principal quantum number specifies the radial variation of this probability.

2. The *angular momentum quantum number*, l, has a much smaller effect upon the total energy of the electron state. It determines the angular variation of this probability. l can take the values 0, 1, 2, 3, . . . , $(n - 1)$.

3. The *magnetic quantum number*, m, has, at most, only a small effect on the electron state energy. It is so named because it determines the component of angular momentum along a particular direction which is of importance when the atom is in a magnetic field. The values of m are restricted to $l, (l - 1), (l - 2), \ldots, -(l - 1), -l$.

4. The *spin quantum number*, s, was first suggested empirically by Uhlenbeck and Goudsmit in 1925. This number can only take the values $+\frac{1}{2}$ or $-\frac{1}{2}$, and it was introduced to explain certain features in atomic spectra by assuming that the electron rotates on its axis about its own center of mass. Later work by Dirac showed that the quantum number s could be introduced as the solution to a particular type of wave equation. The two values $\pm\frac{1}{2}$ denote that the electron can rotate in either direction with $\frac{1}{2}$ quantum unit of angular momentum.

It is not possible to fit the idea of four quantum numbers into an accurate mechanical model of the atom. However, the mechanical model is often retained in mind to give some picture (however inaccurate) of the atom. In

this case, n may be thought of as specifying the major axis of an elliptical orbit; l specifies the minor axis of this orbit; m determines the position of the orbit in space; and s determines the direction in which the electron is spinning on its own axis.

The energy level of the orbit is mainly decided by the principal quantum number, n, and the other quantum numbers have a relatively small influence on the level. However, for many-electron atoms, there are now small differences in energy among electrons having the same value of n, and this explains the presence of many frequencies in the absorption and emission spectra that the Bohr theory was unable to predict.

Figure 1–6 shows diagrammatically the first ten states which can be computed from the solution of the wave equation. The values of the quantum numbers for each of these states can be found by applying the algebraic rules given at the beginning of this section. For example, with $n = 1$, l and m are restricted to zero. The spin quantum number, s, can then take the two values $\pm \frac{1}{2}$, and so there are two states corresponding to $n = 1$. For $n = 2$, there are eight possible states, two with $l = 0$ and six with $l = 1$.

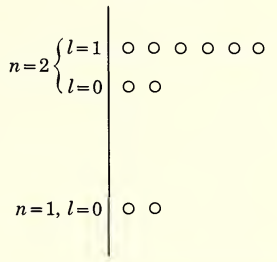


FIG. 1–6. Diagrammatic representation of the first ten energy states.

All electrons which have the same value of n are said to be in the same *shell*. A shell may be further divided into *subshells* where each subshell has a different value of l. Thus, in Fig. 1–6, there are two subshells in the $n = 2$ shell. For the hydrogen atom, the levels shown in the first and second shells are the first ten allowed states where the electron may be found. When the atom is not excited or ionized, the electron is in the $n = 1$ shell. Transitions between the various states correspond to known emission and absorption frequencies.

1–10 The Atomic Table

Unlike the Bohr theory, the quantum mechanical solution of the atom is not limited to hydrogen but can be applied to all atoms. It is possible to compute the electron energy states of the atoms successively from hydrogen up through the atomic table. The element that has two electrons in its atom is helium. Lithium, berylium, boron, and carbon follow with three, four, five, and six electrons respectively. By considering the solution of the Schrödinger wave equation, it is possible to build up an atomic table of the elements as in Table 1–1. Here, the values of n and l are given for the first 36 atoms. The column labeled *atomic number* is the positive charge on the nucleus and is equal to the number of electrons outside the nucleus since the atom as a whole is electrically neutral.

TABLE 1–1. The atomic table for the first 36 elements.

Atomic number	*Element*	*First shell*	*Second shell*		*Third shell*			*Fourth shell*	
		$n = 1$	$n = 2$		$n = 3$			$n = 4$	
Z		$l = 0$	$l = 0$	$l = 1$	$l = 0$	$l = 1$	$l = 2$	$l = 0$	$l = 1$
1	H	1							
2	He	2							
3	Li	2	1						
4	Be	2	2						
5	B	2	2	1					
6	C	2	2	2					
7	N	2	2	3					
8	O	2	2	4					
9	F	2	2	5					
10	Ne	2	2	6					
11	Na	2	2	6	1				
12	Mg	2	2	6	2				
13	Al	2	2	6	2	1			
14	Si	2	2	6	2	2			
15	P	2	2	6	2	3			
16	S	2	2	6	2	4			
17	Cl	2	2	6	2	5			
18	A	2	2	6	2	6			
19	K	2	2	6	2	6		1	
20	Ca	2	2	6	2	6		2	
21	Sc	2	2	6	2	6	1	2	
22	Ti	2	2	6	2	6	2	2	
23	V	2	2	6	2	6	3	2	
24	Cr	2	2	6	2	6	5	1	
25	Mn	2	2	6	2	6	5	2	
26	Fe	2	2	6	2	6	6	2	
27	Co	2	2	6	2	6	7	2	
28	Ni	2	2	6	2	6	8	2	
29	Cu	2	2	6	2	6	10	1	
30	Zn	2	2	6	2	6	10	2	
31	Ga	2	2	6	2	6	10	2	1
32	Ge	2	2	6	2	6	10	2	2
33	As	2	2	6	2	6	10	2	3
34	Se	2	2	6	2	6	10	2	4
35	Br	2	2	6	2	6	10	2	5
36	Kr	2	2	6	2	6	10	2	6

There are two rules which govern how the atomic table is built up. First, in any atom, the electrons will normally occupy the states having the lowest energy. Secondly, no two electrons in an atom can have the same set of four quantum numbers. This second factor was first enunciated by Pauli in 1925 and is known as the *exclusion principle.*

The first ten elements, which have electrons only in the first or second shells, are arranged in the table in the regular manner shown. Only values of n and l are given and the numbers in each column show how many electrons have that value of l (although they will have different values of m and s). From the table it is seen that both levels in the first shell are occupied for helium. Lithium has the first shell filled and an additional electron in an $n = 2$, $l = 0$ state. Berylium has the same filled first shell and two electrons in the $n = 2$, $l = 0$ subshell. This sequence continues up the table until element number ten, neon, is found to have a completely filled second shell.

The third and later shells do not possess the simple regularity shown in the table for the first two shells. It is seen that argon ($Z = 18$) has a closed $l = 1$ subshell but has no electrons corresponding to $l = 2$. Other irregularities occur in the fourth shell. In all cases, however, the lowest empty energy state is filled by the extra electron of the succeeding element.

Inspection of the atomic table shows that the substances having the maximum number of electrons in a shell or subshell are helium ($n = 1$), neon ($n = 2$), argon ($n = 3$), and krypton ($n = 4$). These substances are said to have closed electron shells (or subshells) and they are chemically inert. Their behavior gives a clue to the way in which elements combine together. The closed outer shell indicates chemical stability, and the number of electrons which an electron has missing from a closed shell, or in excess of a closed shell, determines the chemical properties of the material.

1–11 Chemical Valence and Atomic Binding

The number of electrons which an element has in excess of the nearest closed shell is known as a positive valence. The number of electrons below the nearest closed shell is a negative valence. When two substances combine, they do so in such a manner that the combination approaches chemical stability. In a simple way, it may be said that compounds will be built up between elements which complement each other to form closed shells. However, when substances combine, the electron energy levels of both substances are altered, and the energies of the electrons in the outer shells are the most affected. So the simple picture of valence is only approximate, and recourse must be made to the mathematical approach of quantum mechanics to determine the exchange of electrons between the atoms in the general case.

The elements fluorine ($Z = 9$) and chlorine ($Z = 17$) have one electron missing from the closed shell $n = 2$ and $n = 3$ respectively (see Table 1–1) and so they have similar chemical properties. When they combine in chemical

compounds, they try to acquire an electron from another substance to complete the electron shell. Upon consulting the atomic table, it can be seen that lithium ($Z = 3$) and sodium ($Z = 11$), etc., are elements which have an electron in excess of the closed shell that can be used in this process. This class of binding between two atoms is known as *ionic* since one atom of the pair gains an electron and becomes negative and the other one loses an electron and becomes positive. The resulting bond between the two elements is due to the coulomb attractive force between the positive and negative ion.

The two semiconducting elements germanium and silicon take part in a binding process which is different in nature from the ionic mechanism described above. These elements have a valence of ± 4 meaning that four electrons are required to close the shell and also that there are four electrons present in the outer shell (or subshell in the case of germanium). The element carbon ($Z = 6$) also has this property, and all three substances take part in *covalent* binding and, in their solid phases, can exist in the same structural form. This form is characterized by each atom having four near neighbors. The four outer (valence) electrons of each atom require four more electrons to produce a closed shell (or subshell). These four electrons are obtained from the four neighbors. Every valence electron, therefore, may be considered to be shared between two atoms: its parent and one of the neighbors. Thus, every atom shares the eight valence electrons it requires with its neighbors. The way in which atoms of any one of these substances combine together to form a solid crystal is important in semiconductor theory, and covalent binding will be described in the next section in terms of the crystal structure of germanium and silicon.

1–12 Crystals

A *crystal* is a solid piece of material in which atoms are arranged in a regular pattern. The structure of the crystal is determined by the way in which the individual atoms are chemically bound to one another. A *perfect crystal* consists of the repetition of a unit pattern or cell of atoms where all the units are oriented in the same direction with no irregularities. The nearest approach to this, in practice, is a *single crystal* where all the unit cells are arranged in the same direction but some structural discontinuities (defects) are present in the material. These discontinuities may be associated with the presence of a foreign atom or may result from temperature variations when the crystal was being formed. The imperfections in a crystal are extremely important in determining the mechanical and electrical properties of the material. Semiconducting devices are constructed from single crystals which are grown with deliberate additions of chemical impurities (imperfections). Single crystals are rarely found in nature but they may be produced using careful crystal growing techniques. Single crystals of germanium or silicon with minute amounts of impurities are used in the manufacture of transistors, and these

crystals possess particular electrical properties which are discussed later in this chapter. In nature, most crystalline substances occur in *polycrystalline* form where different crystal orientation is found in different parts of the material. Such crystals are generally unsuitable for semiconducting devices.

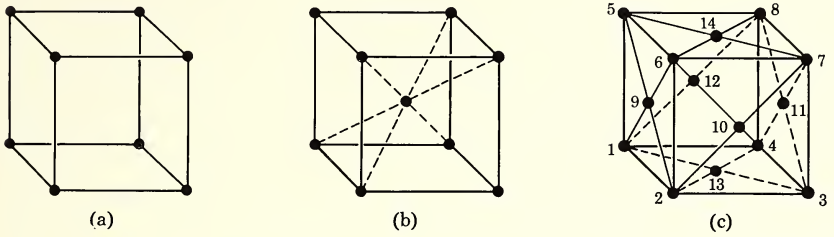


FIG. 1-7. Three types of crystal unit cells.

The germanium and silicon crystal structure is similar to the diamond structure of carbon in that atoms are arranged with covalent bonds to their four nearest neighbors. The unit cell of this type of structure is not simple to understand at first sight as a glance at Fig. 1-8 will show. Before discussing the diamond crystal, consider the three unit cubic cells drawn in Fig. 1-7. In Fig. 1-7(a), the *simple cell* has atoms at each corner. By considering a

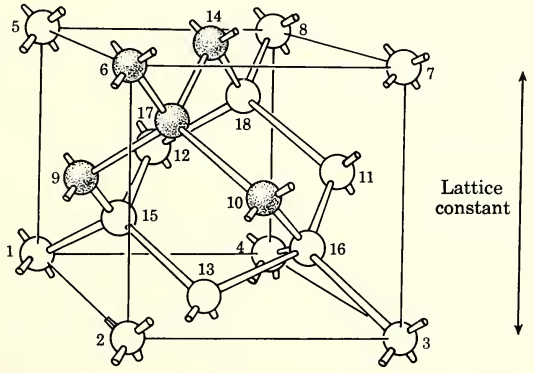


FIG. 1-8. Crystal structure of diamond, silicon, and germanium. A model of the diamond lattice. Each sphere represents a nucleus surrounded by its inner shells. The spokes represent covalent bonds.

number of these cells placed together, it will be seen that every atom has six nearest neighbors spaced apart by the side of the cube. Figure 1-7(b) shows a more complex arrangement which is known as a *body-centered cubic cell.* Here there is one additional atom in the middle of the cube, and this atom has eight nearest neighbors. Finally, the face-centered cube is shown in Fig.

1–7(c). This structure will be shown to be a component of the diamond crystal. For the same size of unit cell, there are more atoms in the face-centered cube than in the body-centered cell and every atom in the face-centered cell will be seen to have twelve nearest neighbors.

The atoms on the diamond structure in Fig. 1–8 have been numbered for identification purposes. Atoms 1 through 8 are at the corners of a face-centered cube. Atoms 9 through 14 are on the six faces of the cube. Four atoms, numbers 15, 16, 17, and 18 remain. These are the atoms of another, similar, face-centered cubic cell which is spaced one quarter of the cube diagonal away from the first unit cell (the diagonal is formed by a line from 1 to 7). Atoms 15, 16, 17, and 18 are the only atoms of this new cell which are contained within the limits of the first cube, and they correspond to numbers 1, 13, 9, and 12 of the first cell. The remaining ten atoms are outside the first cell and are not shown on the diagram. As an aid in picturing the three-dimensional structure, consider the five horizontal layers of atoms in the cube. The bottom layer consists of atoms 1, 2, 3, 4, and 13 on the base of the cube. The second layer which is one quarter of the way up the cube contains the added atoms numbers 15 and 16. Halfway up the cell are the four original face atoms 9, 10, 11, and 12, and three quarters of the way up are added atoms 17 and 18. The top face contains numbers 5, 6, 7, 8, and 14. The distance between atoms 3 and 7 is known as the *lattice constant* for the crystal. The lattice constants for the three materials are: diamond, 3.56 A; silicon, 5.43 A; germanium, 5.66 A.

Consider now atom 17 in relation to its nearest neighbors, atoms 9, 10, 6, and 14 (shown shaded). These five atoms share electrons in covalent bonds. Any other atom in the cell can be shown to have four nearest neighbors although not all of these neighbors are on the diagram.

1–13 Electron Energy Levels in Crystals: Band Theory

The electron energy level diagrams that have been given for single, free atoms will no longer apply to the same atom in a crystal. The energy levels of the inner electrons will be changed only a small amount since the inner (closed shell) electrons are not affected very much by the presence of neighboring atoms. However, the levels of the outer (valence) electrons are greatly affected since these are the electrons which are participating in chemical bonds. Once more, recourse must be made to quantum mechanics in order to determine the new energy levels since classical theory is inadequate to explain the binding process.

There is a simple analogy which may be drawn between the energy of an electron in an outer orbit and the frequency of a tuned LC circuit. In an atom that is isolated from its neighbors, an electron has been shown to possess a certain discrete energy level. Similarly, an isolated tuned circuit possesses a

single resonant frequency of value

$$f_0 = \frac{1}{2\pi\sqrt{LC}}$$

when the circuit resistance is neglected. The analogy between the energy of the electron state and the frequency of the resonant circuit may be extended by considering the effects of coupling in both cases. Figure 1–9 illustrates the effect of coupling several identical tuned circuits together. A sharply tuned resonance curve is obtained for the isolated tuned circuit of Fig. 1–9(a). When two similar circuits are coupled together, a resonance curve having two peaks as in Fig. 1–9(b) is obtained. The introduction of further coupled circuits results in the broad resonance curve as shown in Fig. 1–9(c). The width of the resonance curve depends only on the degree of coupling between the circuits and is independent of the number of circuits which are coupled together.

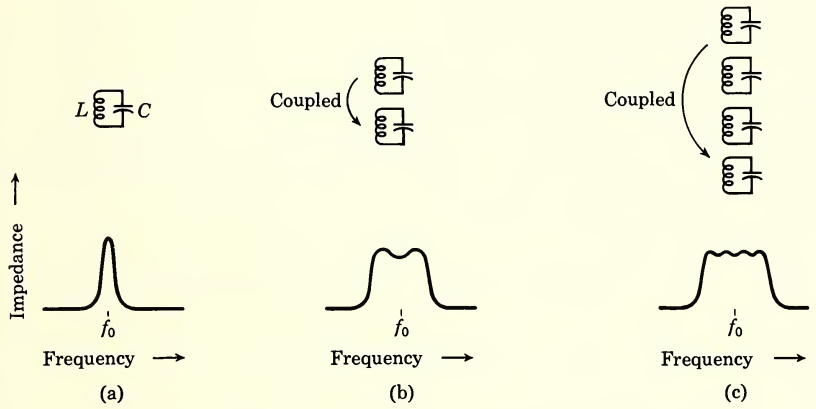


FIG. 1–9. The effect of coupling upon the resonance curve of tuned circuits.

Coupling between the outer electrons of neighboring atoms in a crystal has an effect upon the electron energy levels that is analogous to the broadening of the resonance curve of coupled tuned circuits. It is found that there is now a *band* of energies instead of a discrete energy level as in the single atom. The energy of the outer electrons of any particular atom in the crystal must lie within this band, and it is not possible to specify an individual level any more than it is possible to specify the individual resonant frequency of one of a large number of coupled circuits. The width of the energy band depends upon the coupling between the outer electrons of the atom, that is, upon the distance between atoms in the crystal. If the interatomic distance is decreased, the coupling increases, and the band width increases as shown in Fig. 1–10. This is a theoretical curve as it is not possible to vary continuously the

interatomic distance in a crystal, but it is introduced here as an aid to the understanding of the band theory of electron energies in crystals.

In carbon, there are two electrons in the $n = 1$ shell and four in the $n = 2$ shell. However, there are eight energy states in the complete $n = 2$ shell. In the isolated, free atom which is shown on the right-hand side of Fig. 1–10, six electron states have been shown filled, and the remaining four states in the $n = 2$ shell are empty. As the interatomic distance is decreased, the discrete levels broaden out into bands, and eventually the bands corresponding to the two subshells of $n = 2$ overlap as shown in the center of the diagram. As the interatomic distance is further decreased, a split occurs between the upper

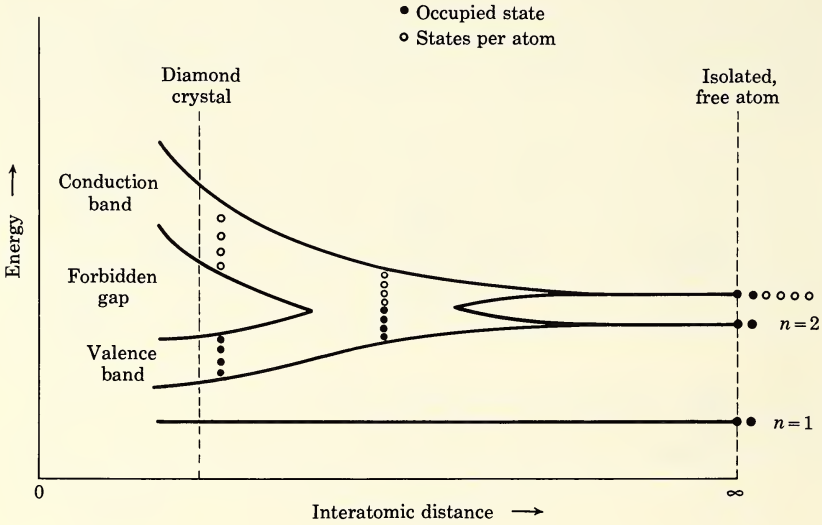


FIG. 1–10. The energy bands in carbon as a function of the interatomic distance (theoretical and not to scale).

four states and the lower four states, and at the point marked "diamond crystal," the two bands are widely separated. The electron level corresponding to $n = 1$ hardly changes as the interatomic distance is decreased.

Figure 1–11 shows an *energy band diagram* for a diamond crystal plotted in the same way as the energy level diagrams that have been discussed previously. This diagram is a section through the curve of Fig. 1–10 on the line marked "diamond crystal" where the energy band for $n = 2$ has split into two parts. The upper part is known as the *conduction band*, and the lower part is the *valence* band. In between the two, there is a region where no electron energy levels are permitted, and this region is known as the *forbidden gap*. At the bottom of the diagram is shown the very narrow band corresponding to $n = 1$. All the bands are shown bounded by straight horizontal lines since the crystal is assumed to be uniform.

An atom of carbon has six electrons and so the lowest six levels per atom will be filled. Thus in Fig. 1–11, the two levels corresponding to $n = 1$ and the four levels per atom in the valence band are filled, leaving the conduction band completely empty. In practice, this is only approximately correct since we have neglected the effect of any heat energy which will be imparted to the electrons in the crystal at any given temperature. If this energy is so small that it has no effect on the distribution of electrons in the levels, conditions are exactly as shown in Fig. 1–11.

The reason for the names of the valence and conduction bands now becomes clear. The four outer electrons per atom, which are in the valence band, are the four electrons that determine the valence of carbon and that are participating in covalent bonds. These electrons, which are being shared with neighboring atoms, are closely bound to the nuclei and so their energy is lower

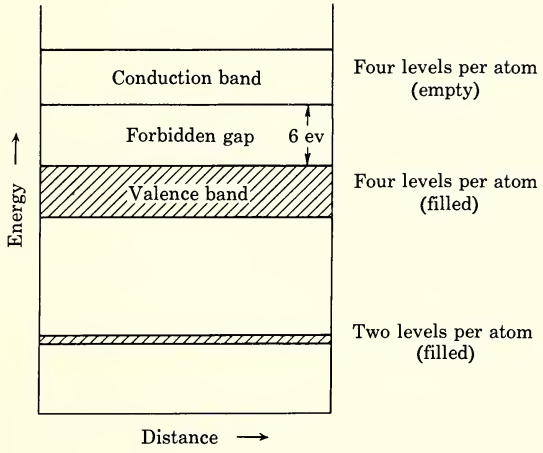


FIG. 1–11. Energy band diagram for a diamond crystal.

(more negative) than the energy of the corresponding states of free atoms (Fig. 1–10). If one electron is removed from a covalent bond in a diamond crystal, it will have acquired sufficient energy to move into an energy state above that of the bound electron. This may be represented on Fig. 1–11 as the passage of an electron from the valence band to the conduction band. The minimum energy that must be imparted to the electron to remove it from the covalent bond is the energy difference between the top of the valence band and the bottom of the conduction band, that is, the width of the forbidden gap (6 ev in a diamond crystal). When the electron is in the conduction band (removed from a covalent bond) it can move from atom to atom since the conduction band is empty. An electron, once removed from the valence band to the conduction band, is available to move in any electric field that may be applied across the material.

1–14 Insulators

Figure 1–11, although applying specifically to the diamond crystal, is also a typical energy band diagram of an insulator. The requirements for a crystalline substance to be an insulator at room temperature are that the valence band must be full, the conduction band must be normally empty, and the forbidden gap must be greater than about one electron volt. Under these conditions, only a minute fraction of the total number of electrons in the material can gain enough energy to cross the forbidden gap and reach the conduction band. It will be shown later that the average energy possessed by an electron at room temperature (300° K) is only 26 millielectron volts

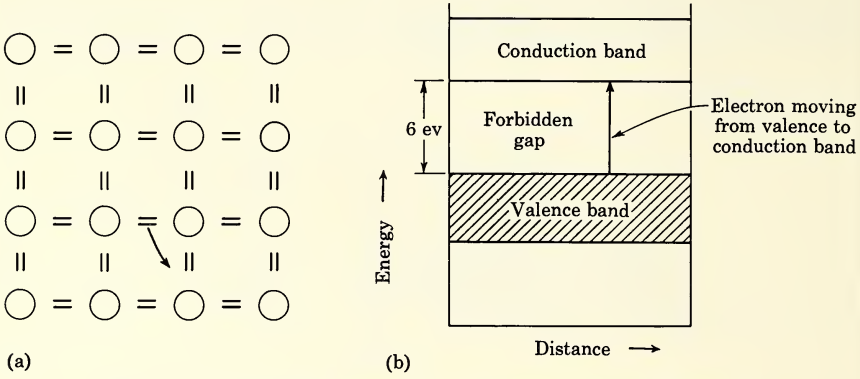


FIG. 1–12. Simplified picture of an electron breaking from a covalent bond in a diamond crystal.

whereas it requires 6 ev of energy to move an electron from the valence to the conduction band in diamond. When an electric field is applied across a crystal, there is a flow of electrons in the conduction band. Diamond, with very few electrons in the conduction band, is therefore an extremely good insulator. If the temperature of the material is increased, the average energy of the electrons rises, and the number of electrons with enough energy to cross the gap also increases. So the conductivity of a good insulator, such as diamond, increases with temperature.

Figure 1–12(a) shows a two-dimensional, simplified picture of a diamond crystal where the electrons that are held in covalent bonds are indicated by short lines. The arrow on this diagram indicates that one electron is about to break away from its bond, and it will then be free to move about in the crystal. In the energy band diagram of Fig. 1–12(b), this process is indicated by the arrow from the valence band to the conduction band. Once in the conduction band, the electron is free from the covalent binding force and can move freely in the crystal. In Fig. 1–12(a), the two inner electrons are not shown since they play no part in the conduction process; the energy band corresponding to these two electrons also has been omitted in Fig. 1–12(b).

1–15 Conductors

In a conductor, the outer electrons are only loosely bound to the parent nucleus, and the permissible electron levels are as indicated in Fig. 1–13. In this diagram, only the highest energy band is shown. This may be a single band, or it may be the result of the overlapping of two energy bands. However it is formed, it has more permitted electron levels than there are available electrons, and so there are empty levels immediately above the highest filled levels. Consider an electron which has an energy corresponding to the top of the filled portion of the band. If this electron gains a small amount of energy from any source, it can move up into the empty part of the band. It is then free to move about in the crystal because the energy states at this level

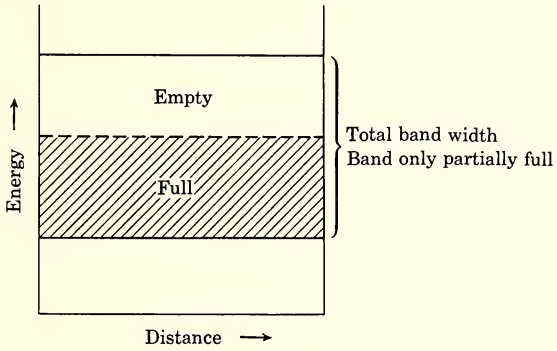


FIG. 1–13. Band structure of a metal.

are not occupied. Since there is no forbidden gap in the energy band just above the filled levels, electrons are always free to move through the crystal under the influence of an electric field. A material of this type is a *conductor*.

1–16 Semiconductors

A semiconductor has the energy band structure of an insulator. However, an insulator has a forbidden gap which is so wide that very few electrons can cross it at room temperature, whereas a semiconductor possesses a narrow gap which allows a considerable amount of conduction at room temperature. Germanium and silicon are the two most used semiconducting elements, and the widths of the forbidden gap in these materials are 0.76 and 1.1 ev respectively. These two materials have a resistivity intermediate between a conductor (such as copper) and an insulator (such as diamond).

When the conductivity in a crystalline semiconductor is due solely to the breaking of covalent bonds, the substance is said to be an *intrinsic semiconductor*. The intrinsic effect is often masked by conduction because of the presence of impurity atoms in the crystal. As explained later in this chapter,

certain atoms introduced into the crystal produce localized energy states which are in the forbidden energy gap, and these may result in electrical conduction. A material where this latter mechanism is dominant is known as an *extrinsic semiconductor.*

I–17 Electrons and Positive Holes

When an electron moves from the valence band into the conduction band of a semiconductor, it leaves behind an unfilled electron state. The absence of an electron in the valence band is called a *positive hole.* This expressive name indicates that a negatively charged electron is missing from the otherwise full band. There are several effects in the crystal that can be attributed to the absence of an electron in the valence band. The positive hole is, therefore,

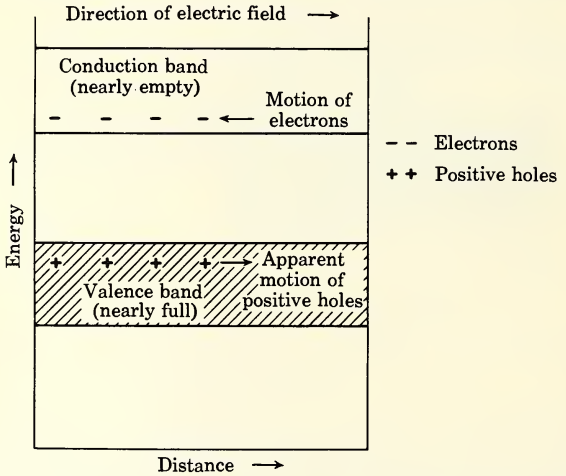


FIG. I–14. Electrical conduction by electrons and positive holes.

endowed with properties that lead to these effects occurring in the crystal. For convenience, the positive hole is regarded as the active particle in the valence band in the same way that the electron is considered to be active in the conduction band. As shown later, this leads to the concept that conduction in semiconductors is caused not only by the motion of electrons in the conduction band but also by positive holes in the valence band.

Before proceeding further, it should be stated that the properties of an electron in a crystal are different from those of a free electron. Forces of attraction and repulsion act upon an electron in a crystal because of the close proximity of the atomic nuclei and of other electrons. In addition, an electron in the conduction band is subjected to different conditions from an electron in the valence band. Thus, the properties assigned to a positive hole (such as mass, mobility, etc.) will not be the same as those of an electron in the conduction band.

The mechanism of *positive hole conduction* can be seen by reference to Fig. 1–14. At 0° K, the valence band is completely full, and the conduction band is empty because no thermal energy is available in the crystal. At any other temperature, as shown in Fig. 1–14, there will be a few electrons which gain enough energy to traverse the forbidden gap to the conduction band and leave positive holes in the valence band. Four such electrons and positive holes are shown. If an electric field, $\mathscr{E}$, is applied across the material such that the left-hand side is positive, electrons in the conduction band, being free to move, will travel from right to left causing an electronic current to flow.

If every electron energy state in the valence band were occupied, there would be no possibility of electrical conduction taking place in this band since a strong electron bond would have to be broken to allow electron motion. However, the absence of four electrons provides four energy states into which neighboring electrons can move under the influence of the electric

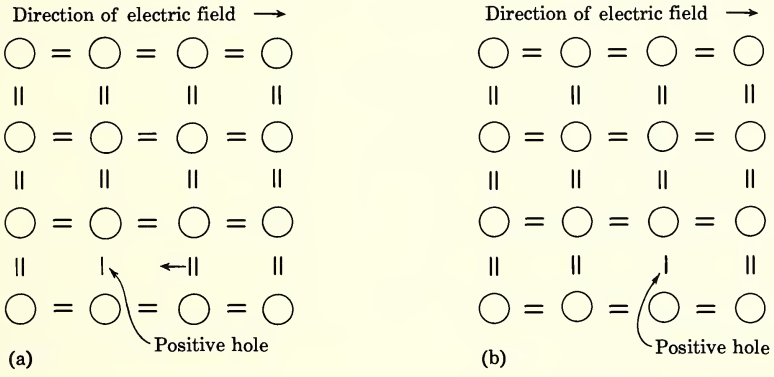


FIG. 1–15. Apparent motion of positive hole in electric field.

field. These electrons require little extra energy to make this transition since the state into which they are moving is in the same energy band.

The motion of a positive hole in a two-dimensional simplified version of a semiconductor crystal is shown in Fig. 1–15. The electric field is in the same direction as in the previous figure. One positive hole is shown in Fig. 1–15(a), and the electric field aids another electron to move into the hole. Part (b) of the figure shows the crystal after the event has taken place. The positive hole has apparently moved in the direction of the electric field, and this motion constitutes a passage of current. This positive hole current takes place exclusively in the valence band since only the closely bound covalent electron energy states are involved. There is also an electron current which occurs in the conduction band and is not shown in Fig. 1–15. This is the result of the drift in the electric field of the electron originally excited into the conduction band.

To distinguish between the two forms of current which flow in a semiconductor it is usual to refer to the motion of electrons in the conduction band and to that of positive holes in the valence band. Yet, it should be remembered that electrons are flowing in both cases, and it is only a matter of convenience to label one a "positive hole" current.

1–18 Mobility

When charged particles in a crystal move under the influence of an electric field, they attain a velocity which is proportional to the value of the field. Thus,

$$v = \mu \mathscr{E} \tag{1–17}$$

where v is the drift velocity in cm sec^{-1}, and $\mathscr{E}$ is the electric field in volts cm^{-1}. The quantity μ is known as the mobility of the particles and is measured in cm^2 $volt^{-1}$ sec^{-1}.

From Eq. (1–17), electrons apparently drift in a given electric field with a constant velocity. However, this velocity is an average value, and the electrons do not, in fact, move across the crystal at a uniform speed. Each electron accelerates in the field and, after traveling a short distance, suffers a collision. At each collision, the electron loses most of the velocity it has gained, and then it starts to accelerate once more under the influence of the field. This process is repeated, and when all the mobile electrons are considered, the average velocity is given by Eq. (1–17).

The value of the mobility in semiconductors depends upon temperature, purity of the material, and on whether electron or positive hole motion is taking place. It is usual to assign the symbol μ_n to the mobility of electrons in the conduction band and μ_p to positive holes in the valence band. For silicon and germanium, μ_n is higher than μ_p because electrons in the conduction band are more free of the attractive forces exerted by the nuclei than are positive holes which move only as a result of the slower drift of electrons in covalent bonds.

TABLE 1–2. Properties of germanium and silicon.

Property	*Germanium*	*Silicon*	*Units*
Atomic number	32	14	
Atomic weight	72.6	28.1	
Density	5.32	2.40	g cm^{-3}
Relative permittivity	16	12	
Gap energy	0.72	1.1	ev
n_i at 300° K	2.5×10^{13}	$\sim 1.6 \times 10^{10}$	cm^{-3}
Intrinsic resistivity at 300° K	47	~230,000	ohm cm
Mobility μ_n at 300° K	3900	1500	cm^2 $volt^{-1}$ sec^{-1}
Mobility μ_p at 300° K	1900	500	cm^2 $volt^{-1}$ sec^{-1}
Diffusion constant D_n at 300° K	99	39	cm^2 sec^{-1}
Diffusion constant D_p at 300° K	49	13	cm^2 sec^{-1}

The values of mobility in Table 1–2 apply to very pure semiconducting material. This is known as *intrinsic* material. It will be shown in Sec. 1–19 that the resistivity of a semiconductor is decreased by the presence of a very small amount of certain types of impurity atoms. If these impurities are present, the mobility also decreases below its intrinsic value. This is due to the increased number of collisions which a charge carrier makes with the added impurity atoms. Curves showing the variation in mobility of electrons and holes in germanium and silicon as a function of the density of added impurity atoms are given in Sec. 1–19.

Consider a semiconductor having an electron density in the conduction band given by n per unit volume. The charge which is available for conduction per unit volume is ne and this charge will move with a velocity v in field $\mathscr{E}$. The current density is given by

$$J_n = nev \tag{1–18}$$

and the conductivity of the material due to electrons in the conduction band only is denoted by σ_n where

$$\sigma_n = \frac{J_n}{\mathscr{E}} = \frac{nev}{\mathscr{E}} = \frac{ne\mu_n\mathscr{E}}{\mathscr{E}} = ne\mu_n \tag{1–19}$$

where the subscript n applies to effects associated with electrons in the conduction band.

If p is the positive hole density in the valence band and μ_p is the positive hole mobility, it can be shown by similar reasoning that

$$\sigma_p = pe\mu_p \tag{1–20}$$

and the total conductivity is

$$\sigma = \sigma_p + \sigma_n = e(p\mu_p + n\mu_n) \tag{1–21}$$

For an intrinsic semiconductor, such as pure germanium or silicon, the electrons in the conduction band have all been removed from the previously full valence band and so

$$p = n = n_i \tag{1–22}$$

where n_i is the intrinsic density of electrons or holes in the material and so

$$\sigma = n_i e(\mu_p + \mu_n) \tag{1–23}$$

1–19 *p*- and *n*-type Semiconductors

In the previous discussion, it was noted that small amounts of certain impurities greatly affect the conductivity of an otherwise pure semiconducting crystal. This will now be explained by reference to the two separate contributions which electron and positive hole motion make to the total conductivity of the material.

Figure 1–16(a) shows the two-dimensional representation of a crystal lattice containing only atoms of an element which has four outer electrons available for covalent binding. Substances of this type (e.g., silicon and germanium) are termed *group 4 materials* and will form a diamond-type

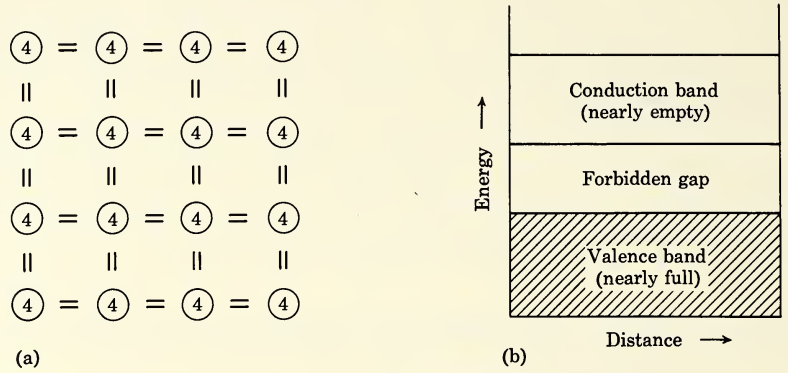


FIG. 1–16. (a) Two-dimensional picture of the crystal structure of a group 4 material. (b) Energy bands of the group 4 crystal.

crystal structure. The energy bands for this crystal are given in Fig. 1–16(b). Now assume that a few atoms of a substance having five outer electrons are introduced into the crystal. This is a *group 5 substance* such as phosphorus, arsenic, or antimony (see Table 1–1), and one atom of this new material is

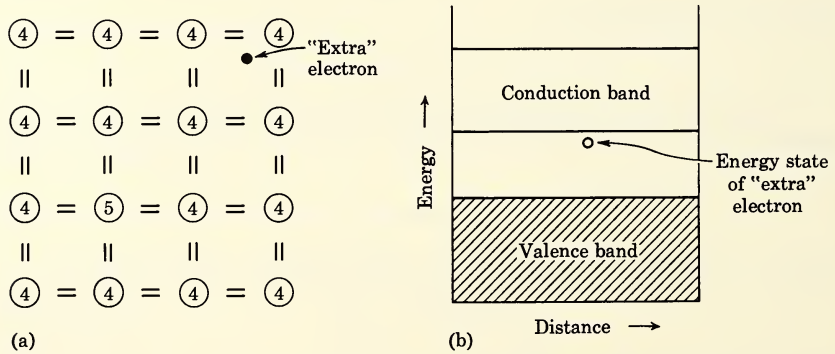


FIG. 1–17. Addition of one atom of group 5 material to a group 4 crystal.

shown in the crystal lattice in Fig. 1–17. Four of the five outer electrons of the group 5 atom will fit into the covalent bonds and will be shared with the four neighboring atoms. The fifth outer electron, however, does not take part in this binding process and is comparatively free to move. It is subject to a force

of attraction towards its parent nucleus, but this force is much smaller than the covalent binding forces represented by energies in the valence band. The energy state corresponding to this fifth electron is shown just below the conduction band in Fig. 1–17 and *is in the forbidden gap* of the pure group 4 substance.

A rough calculation of the force of attraction between the fifth outer electron and its parent atom can be made by referring back to the simple Bohr theory of the atom. If the fifth electron were not present, the remainder of the atom would have a net charge of $+e$. This may be considered as the nuclear charge of a new, Bohr-like atom with one electron in orbit around it. However, Bohr's analysis, as exemplified by Eq. (1–12), was applied to the case where the electron was moving in free space, whereas here the electron is in a crystalline material of relative permittivity, ε_r, greater than unity. Equation (1–12) is, therefore, modified to

$$E_n' = -\frac{me^4}{8n^2h^2\varepsilon_0^2\varepsilon_r^2} \tag{1–24}$$

where E_n' is the energy of the electron in the orbit number n. From Eqs. (1–13) and (1–14), the energy of the ground state of hydrogen ($n = 1$) is

$$E_1 = -\frac{me^4}{8h^2\varepsilon_0^2} = 13.60 \text{ ev} \tag{1–25}$$

Thus, for the electron in the crystal

$$E_1' = \frac{E_1}{\varepsilon_r^2} = \frac{13.60}{\varepsilon_r^2} \text{ ev} \tag{1–26}$$

In a germanium crystal, $\varepsilon_r = 16$ and so

$$E_1' = \frac{13.60}{16^2} = 0.053 \text{ ev} \tag{1–27}$$

It is necessary to justify the assumption that the electron is far enough away from its parent atom that it is surrounded by germanium atoms. From Eq. (1–6), the radius of the first orbit becomes

$$r_1' = \frac{h^2\varepsilon_0\varepsilon_r}{\pi me^4} = \varepsilon_r r_B \tag{1–28}$$

For germanium,

$$r_1' = 16r_B = 8.5 \text{ A} \tag{1–29}$$

This value for the radius of the orbit of the fifth outer electron is greater than the lattice constant of germanium (5.66 A) and so the electron orbit is wide enough to carry it well away from its parent atom and into the germanium crystal (see Fig. 1–17).

TABLE 1–3. Common donor and acceptor elements and their energy levels in germanium and silicon.

Element	*Type of impurity*	*Semiconductor* Ge	Si	*Position*
Phosphorous (P)	Donor	0.012	0.044	In ev below the bottom of the conduction band
Arsenic (As)	Donor	0.013	0.049	
Antimony (Sb)	Donor	0.010	0.039	
Boron (B)	Acceptor	0.010	0.045	In ev above the top of the valence band
Aluminium (Al)	Acceptor	0.010	0.057	
Gallium (Ga)	Acceptor	0.011	0.065	
Indium (In)	Acceptor	0.011	0.160	

The Bohr theory presents a very simple picture of the atom, and the accuracy of the calculation is poor. Measurements have shown that the energy level of the fifth electron of a group 5 atom in a germanium crystal is about

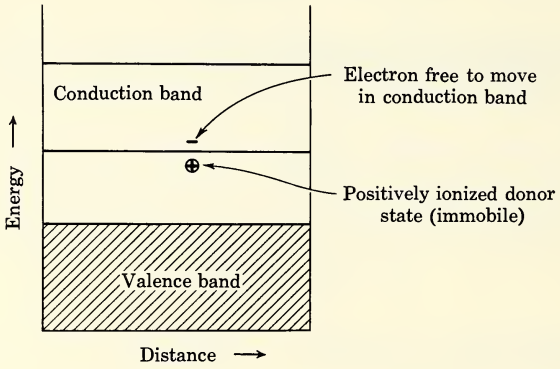


FIG. 1–18. Representation on an energy band diagram of an electron being separated from a donor state.

0.01 ev. For silicon ($\varepsilon_r = 12$), the measured values are commonly about 0.04 ev. Slightly different values are found for different added elements (see "donor" in Table 1–3).

Since it requires such a small energy to remove the electron from the group 5 atom, this energy can be supplied from thermal agitation in the crystal. The electron is usually separated from its parent atom at room temperatures and Fig. 1–18 shows how this process is represented on the energy band diagram. The electron has been removed from the parent atom leaving behind a positively ionized (unfilled) electron bond, and it is free to move under the influence of an electric field. However, *there is no corresponding positive hole in the valence band*, and so only the electron motion in the conduction band will contribute towards the conductivity. The extra energy state associated with the group 5 material is known as a *donor state*. When

the electron is removed, the donor state becomes positively ionized, but the positive charge is fixed in the crystal lattice and cannot move under the influence of an electric field.

In intrinsic semiconductors where no impurity atoms are present, the number of electrons in the conduction band at room temperatures is a minute fraction of the total number of electrons in the crystal. Since each group 5 atom donates one electron to the conduction band, *adding a very small fraction of impurity atoms to the number of group 4 atoms in the crystal increases the conductivity appreciably.* As an example, a ratio of 1 atom of phosphorus to 10^8 atoms of germanium alters the conductivity of germanium at 27° C by a ratio of 12:1.

When the conductivity of a semiconductor is determined largely by the added group 5 atoms, the material is known as *n* (for negative) type. It must be emphasized that the material as a whole is neutral since the added group 5 atoms have as many positive charges in the nucleus as they have

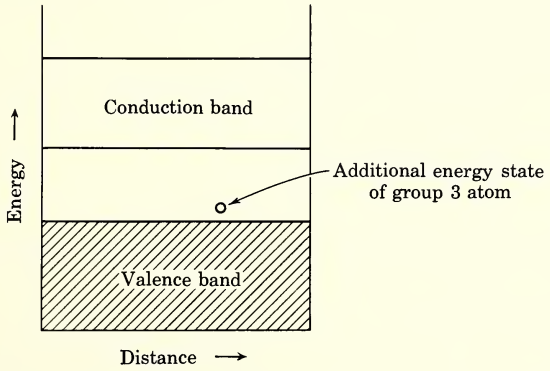


FIG. 1–19. Position of the unfilled energy state of group 3 atom.

negative electrons. The term *n-type* is used to indicate that the major part of the current is carried by electrons in the conduction band. Thus, electrons are known as the *majority carriers*, and positive holes are the *minority carriers* in the *n*-type material.

An atom of *group 3 material* (such as boron, aluminum, etc.) has only three electrons which are available to take part in covalent bonds. Thus for every group 3 atom which is introduced into the crystal, there is one bond which is left unsatisfied and this introduces an *unfilled energy state* into the crystal. The position of this state on the energy band diagram is given in Fig. 1–19. When the group 3 atom acquires an additional electron from a nearby atom to satisfy all its covalent bonds it becomes a negative ion. The energy required for this process is small, and so the new energy state is just above the energy of the electron in its previous state, that is, just above the

valence band. By an extension of the previous application of the Bohr theory of the atom this energy may be calculated. Table 1–3 shows the measured values of the energy level for various group 3 atoms present in germanium and silicon crystals (see "acceptor" in the table).

At room temperature, the crystal has sufficient energy from thermal agitation to allow the energy state to be filled by an electron from the valence band. Figure 1–20 illustrates the condition that exists at room temperature. An electron from the valence band (i.e., an electron originally in a covalent bond somewhere in the crystal) gains sufficient energy to move into the state associated with the group 3 atoms. This leaves a positive hole in the valence band which is available for electrical conduction. The negatively ionized

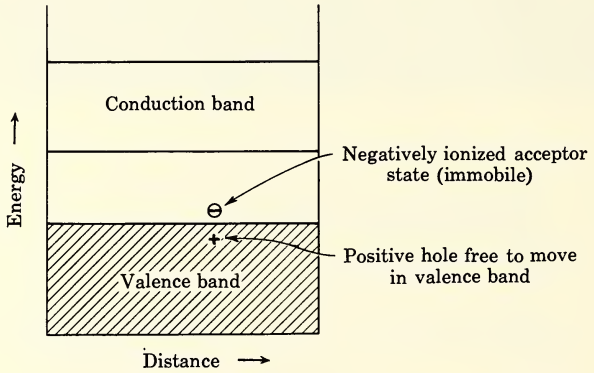


FIG. 1–20. The production of a positive hole in the valence band band of *p*-type material.

filled electron state is fixed in the crystal lattice and makes no contribution to the conduction process. The energy state associated with the group 3 material is called an *acceptor state* since it accepts electrons from the valence band.

The addition of atoms of group 3 material causes a semiconductor to become *p-type*. In this case, positive holes are the *majority carriers*, and electrons are the *minority carriers*. The addition of only a small fraction of impurity atoms increases the conductivity by a large factor (as was the case for the *n*-type material), but now the conductivity is mainly the result of the motion of positive holes in the semiconductor.

We can now distinguish three types of semiconducting material. An *intrinsic* semiconductor is the pure group 4 crystal where the number of holes equals the number of electrons. Material is known as *n-type* when some atoms of group 5 are present in the crystal and there are more electrons than holes. When group 3 atoms are present, the material is *p-type* and holes predominate over electrons.

Other distinguishing features of the three types of material will be pointed out in Chaps. 2, 3, and 4. In particular, we will show that a crystal containing

a *p*-type to *n*-type junction (a *p-n* junction) has the electrical characteristics of a rectifier.

The process of adding either donor or acceptor material to an otherwise pure crystal is referred to as *doping* the crystal. In preparation, the group 4 material is purified to better than a few parts in 10^9, and the required impurity is added before the crystal is formed. If the same proportions of donor and acceptor material are added, the two effects cancel out, and the material becomes intrinsic again. It is the excess of one type of impurity over the other that determines the dominant type of conductivity (see Sec. 2–8).

Figures 1–21 through 1–24 show the measured values of electron and hole mobilities in *n*- and *p*-type germanium and silicon. The curves were obtained by measurement and calculation on semiconductor specimens of known resistivity. In each case, the mobility is plotted against the resistivity which is marked as an open scale at the top of the graph. The bottom scale is the density of impurity atoms in the material. N_d signifies the number of donor atoms per cubic centimeter for *n*-type material, and N_a is the number of acceptor atoms per cubic centimeter for *p*-type material. These curves, therefore, allow us to determine the resistivity when the donor or acceptor density is known by comparing the scales at the top and the bottom of each graph. The equations relating mobility, resistivity, and impurity density are derived in Sec. 2–8.

From Figs. 1–21 and 1–22, which relate to germanium, it can be seen that the addition of as few as 10^{14} atoms of donor or acceptor material per cubic centimeter reduces the resistivity of the specimen appreciably. Since there are 4.42×10^{22} atoms of germanium per cubic centimeter, the impurity atoms form a minute proportion of the total number of atoms in the crystal but their effect on the resistivity is high. For both silicon and germanium, the curves show that the mobility decreases with decreasing resistivity for material of resistivity lower than about 10 ohm cm.

1–20 Generation and Recombination

Because of the vibrational energy of the crystal lattice at room temperatures, covalent bonds are continually being broken, and electrons are being set free. This process is known as the *generation of electron-hole pairs*. The reverse process is also taking place, and electrons are being recaptured by unfilled bonds to produce the *recombination of electron-hole pairs*. Both generation and recombination are random processes which are continuously occurring. When the temperature and external conditions are constant, a balance is set up between generation and recombination, and the resultant hole and electron densities are known as the *thermal equilibrium values*, p_0 and n_0 respectively. Since there are usually more than 10^{10} carriers per cubic centimeter in a typical piece of semiconducting material, the statistical variation in these quantities is small and can be neglected. The symbols, p and n, are retained for the carrier densities under general conditions.

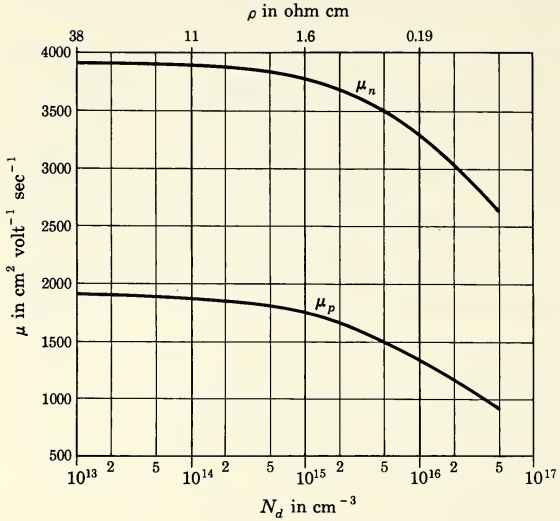


FIG. 1–21. Electron and hole mobilities in *n*-type germanium at 300° K [from M. B. Prince, *Phys. Rev.* **92** (1953), 681; **93** (1954), 1204].

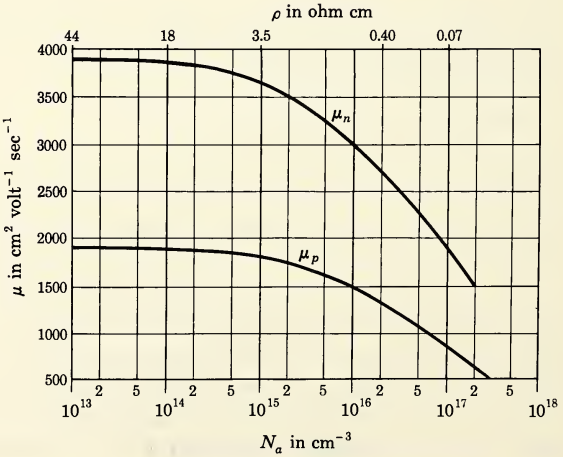


FIG. 1–22. Electron and hole mobilities in *p*-type germanium at 300° K [from M. B. Prince, *Phys. Rev.*, **92** (1953), 681; **93** (1954), 1204].

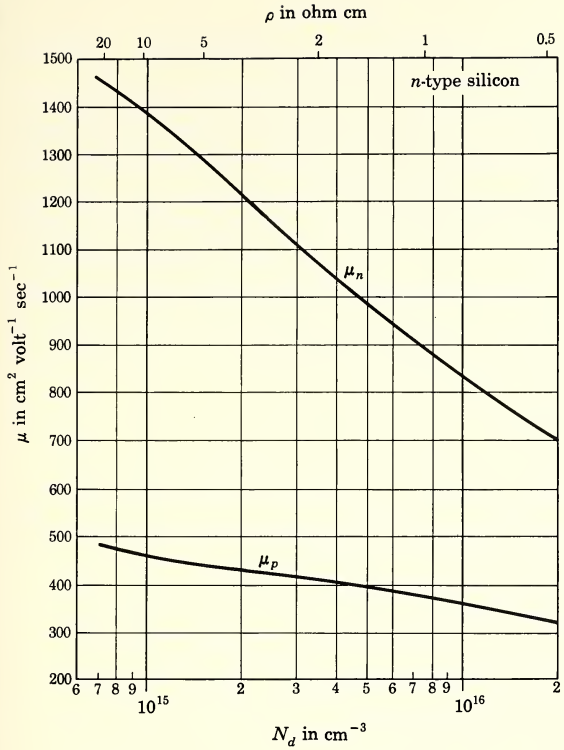


FIG. 1–23. Electron and hole mobilities in *n*-type silicon at 300° K [from M. B. Prince, *Phys. Rev.*, **92** (1953), 681; **93** (1954), 1204].

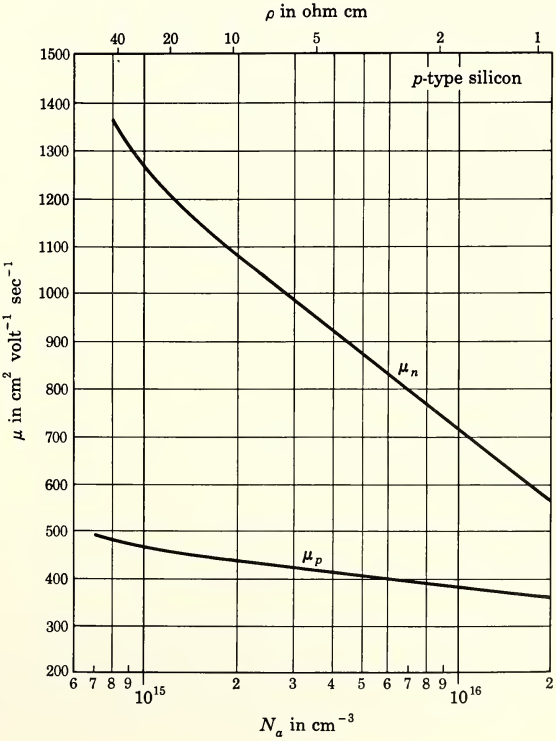


FIG. 1–24. Electron and hole mobilities in *p*-type silicon at 300° K [from M. B. Prince, *Phys. Rev.*, **92** (1953), 681; **93** (1954), 1204].

It has been shown that the production of an electron-hole pair involves the passage of an electron from the valence band to the conduction band. Similarly, recombination is the process whereby an electron moves from the conduction band to the valence band. The direct transition of an electron from the conduction band to the valence band is unlikely, however, since a free electron and a hole will only recombine if momentum is conserved in the system, that is, if the electron and the hole initially have nearly equal and opposite momenta. The probability for this in germanium and silicon is low, and another recombination mechanism must be invoked.

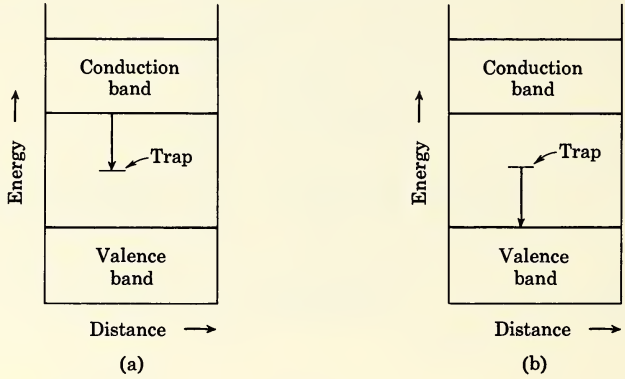


FIG. 1–25. (a) An electron moves from the conduction band into a recombination trap. (b) The electron moves from the trap down into the valence band.

The majority of recombination events are found to take place at the site of an impurity atom or a lattice defect in the crystal. Such a location acts as a *trap* or *recombination center* and is, in effect, a third body which can satisfy the momentum requirements in the electron-hole collision. Figure 1–25 shows a trap which is located approximately in the center of the forbidden gap. The electron moves from the conduction band into the trap level and remains there for a short time before passing down into the valence band. Once an electron is in the trap it is not free to move and does not contribute to the conductivity.

The average time that an electron spends in the conduction band before recombining via a trap is known as the *lifetime*, τ_n. Similarly, τ_p is the average lifetime of holes in the valence band. Electron and hole lifetimes are in the range 10^{-8} to 10^{-3} second for the semiconducting material used in transistors. The higher the conductivity the lower the lifetime since the added group 3 or 5 atoms act as recombination centers. Lifetimes also depend upon the purity of the material (other than the presence of group 3 and 5 atoms), temperature, and shape and surface conditions of the specimen (see Sec. 9–10). In this book, we shall be primarily concerned with *minority carrier lifetimes*, for example, the lifetime of electrons in *p*-type material.

Consider unit volume of a piece of p-type material which has a thermal equilibrium electron and positive hole density n_0 and p_0 respectively. If excess electrons are introduced into the crystal, an equation can be derived to determine the way in which the density decays back to its original value. There is a *generation rate*, g, of electron-hole pairs per unit volume that is dependent upon the temperature for a given material but not upon the number of electrons present. If the electron lifetime is τ_n, and n electrons are present at time t, then the number of electrons recombining per second is n/τ_n. Thus,

$$\frac{dn}{dt} = g - \frac{n}{\tau_n} \tag{1-30}$$

where dn/dt is the rate of change of density with time (a negative quantity in this case). At equilibrium, $dn/dt = 0$ and $n = n_0$, and so

$$g = \frac{n_0}{\tau_n} \tag{1-31}$$

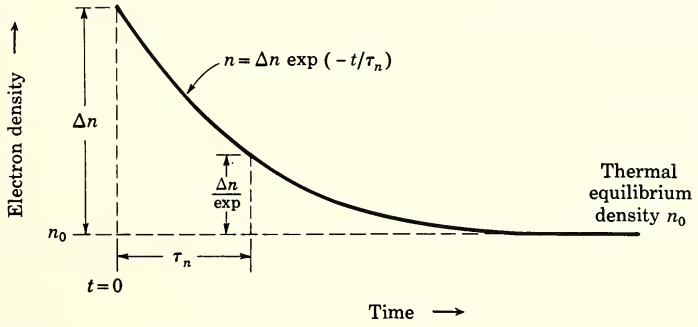


FIG. 1-26. The decay of excess carriers in a semiconductor [Eq. (1-33)].

Equation (1-30) can now be rewritten as

$$\frac{dn}{dt} = \frac{n_0 - n}{\tau_n} \tag{1-32}$$

and the solution is

$$n - n_0 = \Delta n \exp\left(-\frac{t}{\tau_n}\right) \tag{1-33}$$

Here, Δn is the value of $n - n_0$ at $t = 0$, i.e., it is the excess electron density introduced into the material. Equation (1-33) and Fig. 1-26 show that the number of excess electrons in the conduction band decays exponentially with a time constant equal to the lifetime of the electrons in the material.

By a similar argument, it may be shown that the corresponding equation for positive holes is

$$p - p_0 = \Delta p \exp\left(-\frac{t}{\tau_p}\right) \tag{1-34}$$

when excess positive holes, Δp, are introduced into n-type material, where τ_p is the lifetime of positive holes.

BIBLIOGRAPHY

General

Dekker, Adrianus J., *Electrical Engineering Materials*, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1959

Shive, J. N., *Properties, Physics and Design of Semiconductor Devices*, New York: D. Van Nostrand and Company, Inc., 1959

Spangenburg, Karl R., *Fundamentals of Electron Devices*, New York: McGraw-Hill Book Company, Inc., 1957

Sproull, Robert L., *Modern Physics*, New York: John Wiley & Sons, Inc., 1956

van der Ziel, Aldert, *Solid State Physical Electronics*, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1957

Atomic physics

Van Name, F. W., Jr., *Modern Physics*, 2d ed. Englewood Cliffs, N.J.: Prentice-Hall, Inc. 1962

Wehr, M. Russell, and James A. Richards, Jr., *Physics of the Atom*, Reading, Mass.: Addison-Wesley Publishing Company, Inc., 1960

Yarwood, John, *Electricity, Magnetism and Atomic Physics*, Vol. II, *Atomic Physics*, London: University Tutorial Press Ltd, 1958

Crystals

Dekker, Adrianus J., *Solid State Physics*, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1957

Kittel, Charles, *Introduction to Solid State Physics*, 2d ed., New York: John Wiley & Sons, Inc., 1956

Elementary semiconductor theory

Dekker, Adrianus J., *Electrical Engineering Materials*, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1959

DeWitt, David, and Arthur L. Rossoff, *Transistor Electronics*, New York: McGraw-Hill Book Company, Inc., 1957

PROBLEMS

1–1 Write the expressions for the magnitudes of the tangential and angular electron orbital velocities as a function of the principal quantum number in the Bohr theory of the hydrogen atom. Calculate these velocities for $n = 1$.

ans. 2.2×10^6 m sec^{-1} 4.15×10^{16} rad sec^{-1}

1–2 Find the energy of an electron in the first, second, and third orbits of the Bohr atom in joules and electron volts. What are the wavelengths of the photons emitted when electrons make the transition between orbits (a) 1 and 2, (b) 2 and 3, (c) 1 and 3?

1–3 Discuss the fundamental postulates used by Bohr in his theory of the hydrogen atom. Explain the meaning of the terms "excitation" and "ionization".

1–4 Calculate the energy of one quantum of the following radiation: (a) X band radar (9300 Mc/sec), (b) far infrared, (c) blue light, (d) 50 kilovolt X rays.

ans. 3.8×10^{-5} ev, 3×10^{-2} ev, 2.6 ev, 5×10^4 ev

1–5 (a) Electromagnetic radiation of wavelength 949.5 A is projected on to hydrogen gas. Is any of the radiation absorbed? If absorption takes place, what are the initial and final energy states of the hydrogen atom? (b) Repeat the question for radiation of wavelength 2108.3 A.

1–6 Before Bohr's theory of the atom was developed, Balmer gave an empirical relationship between the wavelengths of some of the spectral lines of hydrogen. This may be put in the form

$$\frac{1}{\lambda} = R\left(\frac{1}{4} - \frac{1}{n^2}\right)$$

where λ is the wavelength of a spectral line in the series, n has the value 3, 4, 5, 6, . . . , and R is known as Rydberg's constant. From Bohr's theory of the atom, explain which transitions produce this series of spectral lines and derive a value for R in terms of the fundamental constants.

ans. $R = (me^4/8\varepsilon_0^2 h^3 c)$

1–7 Bohr's theory of the atom applies to any one-electron atom and can, therefore, be applied to singly ionized helium and doubly ionized lithium. What energy is required to remove the remaining electron from these two atoms, respectively?

1–8 What is the de Broglie wavelength in Angstrom units of (a) an electron of energy 10 ev, (b) a proton of energy 1 million-electron volts, (c) a satellite of mass 100 lb orbiting the earth with a velocity of 18,000 mph?

ans. 3.9 A, 2.86×10^{-4} A, 1.82×10^{-29} A

1–9 Explain why the discrete energy levels for a free, single atom will no longer apply to the same atom in a crystal. Distinguish among conductors, insulators, and semiconductors.

1–10 If the lattice constant for Fig. 1–8 is 5.66 A, what is the distance between neighboring atoms?

1–11 A germanium crystal specimen 1 mm cube has a total of 2.5×10^{7} electrons in its conduction band. What electron current flows when there is a field of 6 volts cm^{-1} parallel to one face of the cube?

1–12 One atom of boron is contained inside an otherwise pure germanium crystal. Using the concepts of covalent binding between neighboring atoms in a crystal, in your own words explain the effect of the added boron atom on the electrical properties of the crystal.

2

Electrons and Holes in Semiconductors

The previous chapter presented a descriptive account of electrical conduction in semiconductors. This chapter provides a more exact analysis of the behavior of electrons and holes. This is carried out by using an elementary treatment of Fermi-Dirac statistics.

2–1 The Density of States in an Energy Band

The energy band representation which has been used in Chap. 1 shows that there are electron energy states in both the valence and conduction bands in a semiconductor and that these bands are separated by a forbidden gap. In the valence band of a group 4 semiconductor, it is known that there are four electron states for each atom (Fig. 1–10), and so the total number of states is four times the number of atoms. This makes it possible to establish the total number of states in the valence band per unit volume but gives no information on the number of holes and electrons in the material.

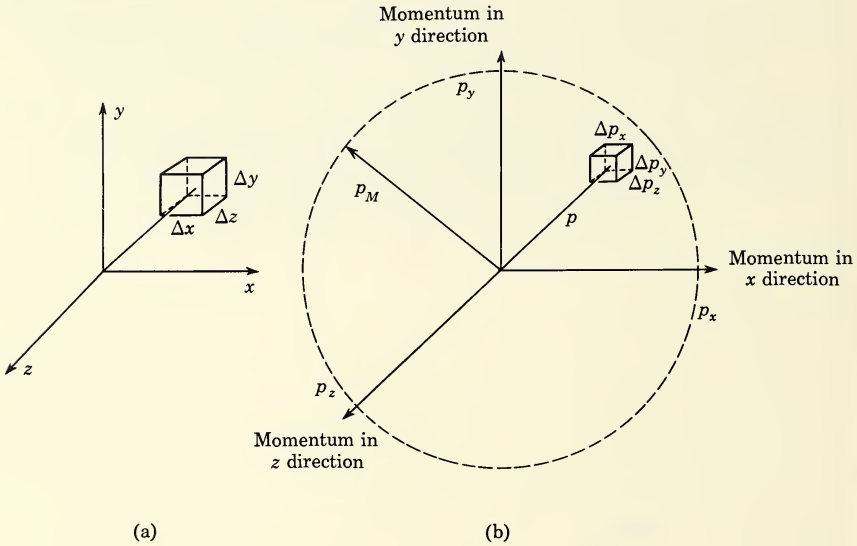


FIG. 2–1. A momentum cell diagram. (a) Cube of metal having sides $\Delta x = \Delta y = \Delta z = 1$. (b) Momentum diagram showing unit cell of dimensions $\Delta p_x = \Delta p_y = \Delta p_z = h$. Dotted circle indicates sphere of radius p_M inside which all cells are occupied when M electrons per unit volume are in the band at 0°K.

An electron-hole pair is produced when an electron acquires sufficient energy to leave the valence band and enter the conduction band. The passage of an electron from the top of the valence band to the bottom of the conduction band requires the expenditure of the least amount of energy of any transition which can be made between the two bands. Transitions will normally be of this type, and so the number of electrons which move between the two bands depends partly on the temperature and partly on the number of electron energy states at the two band edges. It is, therefore, important to learn how the states are distributed with respect to energy in the valence and

conduction bands. We will first consider the simple case of the energy states in the outer (conduction) band of a metal. In Sec. 2–2, this result will be applied to the valence and conduction bands in a semiconductor.

Consider a cube of metal having unit sides in the three coordinate directions as shown in Fig. 2–1(a). From Sec. 1–15, the electrons in the outer energy band of the metal are free to move within the limits of the unit cube. Thus, an electron cannot be located more accurately than to state that it is still in the cube, that is, its position is specified to within Δx, Δy, and Δz in Fig. 2–1(a), where

$$\Delta x = \Delta y = \Delta z = 1 \tag{2–1}$$

The momentum of the electron will next be considered. If p_x, p_y, p_z are the three components of the momentum in the three coordinate directions, we are able to specify the momentum to within Δp_x, Δp_y, and Δp_z by applying *Heisenberg's uncertainty principle* as expressed in Eq. (1–16). Thus,

$$\Delta p_x = \Delta p_y = \Delta p_z = h \tag{2–2}$$

This application of the uncertainty principle is illustrated in Fig. 2–1(b) where the three axes indicate momentum in the three coordinate directions, and the vector, p, represents the magnitude and direction of the momentum of any one electron. The uncertainty principle shows that, for a unit cube of the metal, it is impossible to specify p with any more accuracy than to state that the line representing p must end somewhere within a cell of side h and volume h^3 on the momentum diagram. This cell is known as the *unit momentum cell*, and it can be seen that other cells having the same dimensions but different positions can be drawn to represent the accuracy of specification of the momenta of other electrons in the metal.

This pictorial representation of momentum cells can be extended by applying the Pauli exclusion principle. In Sec. 1–10, it was stated that no two electrons can have the same set of four quantum numbers in an atom and, because of the close coupling between atoms, the exclusion principle applies in the metal crystal as well. Since the spin quantum number s can take only the values $+\frac{1}{2}$ or $-\frac{1}{2}$, the exclusion principle may be restated in the form that only two electrons can have the same three quantum numbers n, l, and m. The principle is now applicable to the momentum diagram since the three quantum numbers determine the momentum p. Thus each unit momentum cell, which indicates the accuracy of specifying one value of p, now accounts for two electrons having opposite spins ($s = +\frac{1}{2}$ and $s = -\frac{1}{2}$).

Consider now that there are M free electrons available in the material. At a temperature of 0° K, there is no disturbing motion in the crystal due to heat energy, and the electrons will occupy cells having the lowest values of momentum (and energy). There is no preferred direction in the crystal, and so $M/2$ cells arranged spherically about the origin as center up to a radius p_M will be occupied as shown in Fig. 2–1(b). The "volume" of each cell is h^3

and so the total "volume" of occupied cells in the momentum diagram is

$$\frac{h^3 M}{2} = \frac{4}{3}\pi p_M^3 \tag{2-3}$$

The term "volume" has been used, and we note that the axes of Fig. 2–1(b) have the dimensions of momentum not length. Figure 2–1(b) is often called a *momentum space diagram.*

It is possible to relate the maximum momentum, p_M, with the position of the corresponding electron energy state, E. On the momentum space diagram, the zero of momentum ($p = 0$) corresponds to the bottom of the

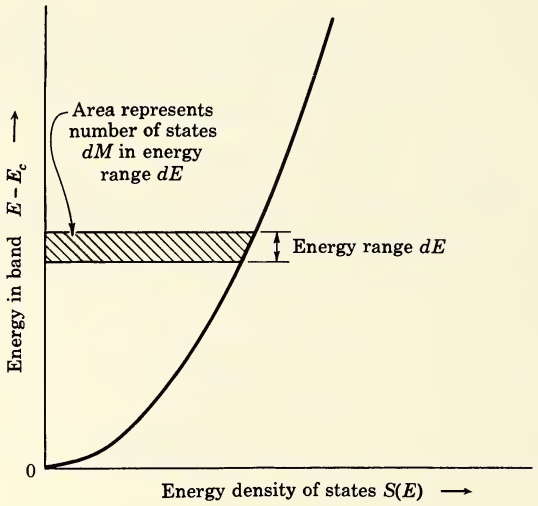


FIG. 2–2. Energy density of states plotted as a function of energy.

energy band which we will call E_c. The energy difference between the levels at E and E_c corresponds to the momentum p_M, or

$$E - E_c = \frac{1}{2} m v_M^2 = \frac{p_M^2}{2m} \tag{2-4}$$

where m is the mass of the electron, and v_M is its velocity. Eliminating p_M between Eqs. (2–3) and (2–4),

$$M = \frac{2^{9/2} m^{3/2} \pi}{3h^3} (E - E_c)^{3/2} \tag{2-5}$$

Equation (2–5) gives the relationship between the total number of electron states which are filled at 0° K and the energy level of the highest state relative to the bottom of the band, $E - E_c$. If M is increased by a small value, dM, the radius of the sphere on the momentum space diagram will increase from p_M to $p_M + dp_M$, and the highest filled energy level will rise from E to $E + dE$.

We now define the *energy density of states* to be $S(E)$ given by the equation

$$S(E) = \frac{dM}{dE} \tag{2-6}$$

where the E in parenthesis indicates that the energy density, S, is a function of energy. The energy density is thus the number of states, dM, present in unit energy range, and indicates the degree of "packing" of the states with respect to energy. This is the most useful way of expressing the information given in Eq. (2-5) since it gives the *numbers of states per unit energy* at any level in the band. By differentiation,

$$S(E) = \frac{dM}{dE} = \frac{2^{7/2}m^{3/2}\pi}{h^3}(E - E_c)^{1/2} \text{ per unit volume} \tag{2-7}$$

Equation (2-7) has been plotted in Fig. 2-2. Energy has been taken as the ordinate in this curve to conform with the energy band diagrams given previously. The shaded area in the figure is equal to the number of states, dM, in energy range dE in accordance with Eq. (2-7). This information, which we have derived for the simple case of a metal, will now be applied to the conduction and valence band in a semiconductor.

2-2 Energy State Densities in Semiconductors and the Effect of Temperature

The energy density of states at the bottom of the conduction band of a semiconductor is given by

$$S(E) = \frac{2^{7/2}m_c^{3/2}\pi}{h^3}(E - E_c)^{1/2} \tag{2-8}$$

This is of similar form to Eq. (2-7) with E_c now the energy at the bottom of the conduction band and m_c as the *effective mass* of the electron in that band. It is necessary to introduce the concept of effective mass to take into account the influences of the neighboring atoms and charges on the motion of an electron in the band. Thus, an electron will respond to an applied force as if it had a mass of m_c (not the free electronic mass, m) when it is in the conduction band. In the valence band, the *effective mass of the positive hole* is written as m_v, and by analogy with the argument of the previous section, we write the energy density of states in the valence band as

$$S(E) = \frac{2^{7/2}m_v^{3/2}\pi}{h^3}(E_v - E)^{1/2} \tag{2-9}$$

where E_v is the energy at the top of the valence band, and E is any energy in the valence band. The values of m_c and m_v are not the same but they will be considered so in this simplified analysis. Note that this introduces a small

error in Eq. (2–22) (see Problem 2–7). The curve of $S(E)$ versus E is given in Fig. 2–3 for the valence and conduction bands of a semiconductor.

The production of hole-electron pairs was shown in Chap. 1 to be the result of the passage of an electron from the valence band to the conduction band. In this chapter, we look at the same phenomenon from a different point of view and consider how many of the electron energy states in the two bands will be occupied. To do this, we introduce the *Fermi factor*, $F(E)$, which is a number that expresses the *probability that a state of a given energy is occupied by an electron.* This number has a value between zero and unity, and is a function of energy and temperature. Zero probability means that the state is empty, and a probability of unity implies that the state is occupied. A probability of 0.5, for instance, means that states at this energy level are 50 per cent occupied. Since this is a statistical process, this probability represents only the *average occupancy rate.* However, the total number of states involved is usually so high that the fluctuation from the statistical average is small.

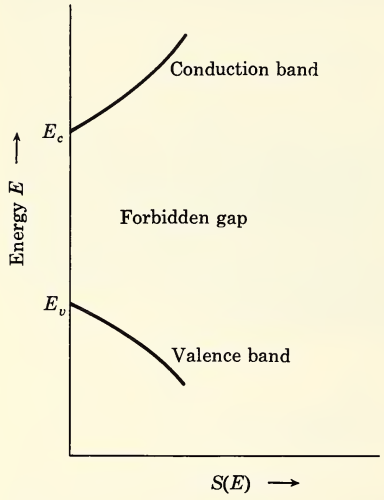


FIG. 2–3. Energy density of states in valence and conduction band.

The Fermi factor, $F(E)$, can be derived by a statistical argument which will not be given here (see the bibliography at the end of this chapter). In essence, it consists of finding the most probable arrangement of electrons in energy states, when occupancy is governed by the Pauli exclusion principle, when there is a given total energy in the system and when the number of electrons is a constant. Under these circumstances the Fermi factor is

$$F(E) = \frac{1}{1 + \exp\left[(E - E_F)/kT\right]} \tag{2–10}$$

where E_F is a particular energy known as the *Fermi level*, k is Boltzmann's constant [1.38×10^{-23} joule $(°K)^{-1}$], and T is the absolute temperature, °K. The Fermi factor is independent of the energy density of states; it is the probability that states are occupied at that level irrespective of the number of states actually present, that is, it is the fractional occupancy of possible states. *Even if there are no states at a particular energy* (as in the forbidden gap of an intrinsic semiconductor), *the Fermi factor still has a definite value.*

Before discussing the Fermi factor further, consider Eq. (2–11).

$$N(E)\,dE = S(E)F(E)\,dE \tag{2–11}$$

Since $S(E)$ is the energy density of states, $S(E)\,dE$ is the total number of available states in the energy range dE. $F(E)$ is the probability that these states are occupied and so the product of the density of available states, $S(E)$, and the probability of their being occupied, $F(E)$, gives the number of occupied states, $N(E)$, in the small energy range considered, dE. Equation (2–11) and this verbal statement of it are important for an understanding of later work in this chapter.

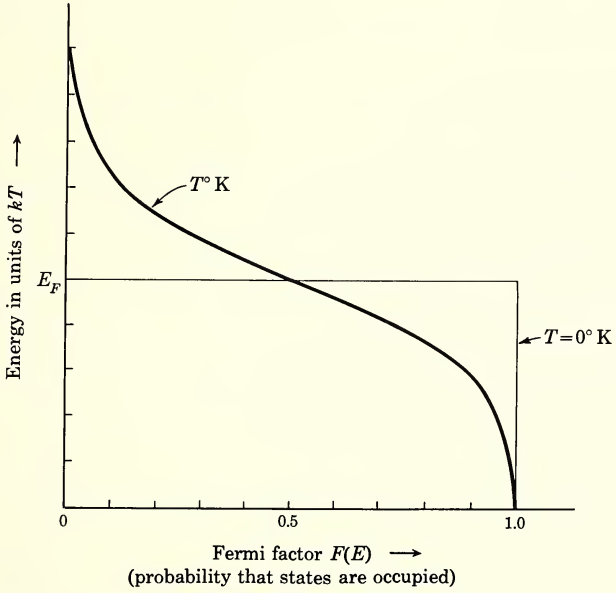


FIG. 2–4. Fermi factor as a function of energy for two different temperatures.

The value of the Fermi factor is determined by the ratio $(E - E_F)/kT$ from Eq. (2–10). The quantity kT has the dimensions of energy and can be specified in electron volts. For example, $kT = 25.8$ millielectron volts at 300° K. The variation of $F(E)$ with energy is shown in Fig. 2–4. As the temperature is increased, the size of the units of energy shown in the figure also increases showing that $F(E)$ is a function of temperature. *The Fermi level, E_F, is a quantity of great significance in semiconductor electronics.* It is a reference level which is an important characteristic of a semiconductor as the following analysis will show.

When the temperature is 0° K, $F(E)$ can have one of two values.

For $E > E_F$, $$F(E) = \frac{1}{1 + \exp(\infty)} = 0$$

For $E < E_F$, $$F(E) = \frac{1}{1 + \exp(-\infty)} = 1$$

Therefore, the value of $F(E)$ is unity for all energies below the Fermi level and zero for all energies above the Fermi level. This corresponds with the case shown in Fig. 2–1 where all the momentum cells were full up to the radius p_M at 0° K. The energy corresponding to p_M must coincide with the Fermi level, and we can say that at 0° K all energy states will be full up to E_F.

A more useful definition of the Fermi level can be obtained by putting $E = E_F$ at any temperature other than 0° K. Then, if $E = E_F$ and $T \neq 0$

$$F(E) = \frac{1}{1 + \exp(0)} = \frac{1}{2}$$

Thus, the Fermi level, E_F, is that energy where the probability of a state being occupied is $\frac{1}{2}$.

Referring to Fig. 2–4, at 0° K all the energy states below the Fermi level are completely occupied whereas above this level they are completely empty. At 0° K there is no heat energy present in the crystal lattice, and so no covalent bonds are being broken. At any other temperature, however, the process of generation and recombination of electron-hole pairs is taking place. Because of this, there is a probability of slightly less than one that all of the states below the Fermi level are occupied, and also a small probability that some of the energy states just above the Fermi level are occupied.

It has been shown that both electrons and holes contribute to the conductivity of semiconductors. In the conduction band, therefore, the probability of electron states being occupied, $F(E)$, is of interest because all electrons in this band are mobile. In the valence band, positive holes are responsible for the passage of charge. Thus, it is the *absence* of electrons which is important in the valence band, and we shall be considering $1 - F(E)$ since this is the probability of a positive hole, rather than a filled state, occurring at a particular energy level in the band.

The concept of the probability of occupancy of an energy state, as represented by $F(E)$ is applicable to all solid-state crystalline materials. Equation (2–10) is true for conductors, insulators, and semiconductors. The position of the Fermi level depends upon the available number of electrons in the material and upon the energy state density curve, $S(E)$. Relative to the Fermi level, however, the probability of an energy state being filled depends only upon the temperature as shown by Eq. (2–10).

The total number of electrons per unit volume in a material is M where

$$M = \int N(E)\,dE = \int S(E)F(E)\,dE \qquad (2\text{–}12)$$

and the limits of integration are from the bottom of the lowest energy band to the top of the highest band concerned. Equation (2–12) can be used to find the position of the Fermi level for any material when the total number of electrons per unit volume, M, is known and where $S(E)$ can be determined.

The value of the Fermi level gives important information about a solid

material since it indicates the highest possible filled electron energy state at 0° K. At any other temperature, as shown in Fig. 2–4 and Eq. (2–10), the value of $F(E)$ drops from 0.95 to 0.05 as the energy changes from $E_F - 3kT$ to $E_F + 3kT$. Below this region, states are almost completely full; above this region, they are almost empty. The Fermi level is in the center of this region.

The ideas presented in this section will be applied to semiconductors to determine the number of electrons in the conduction band, the number of holes in the valence band, and the position of the Fermi level. The case of an intrinsic semiconductor will be considered first.

2–3 Intrinsic Semiconductor

In a specimen of unit volume, the total number of occupied states with energy between E and $E + dE$ is equal to the product of the density of the

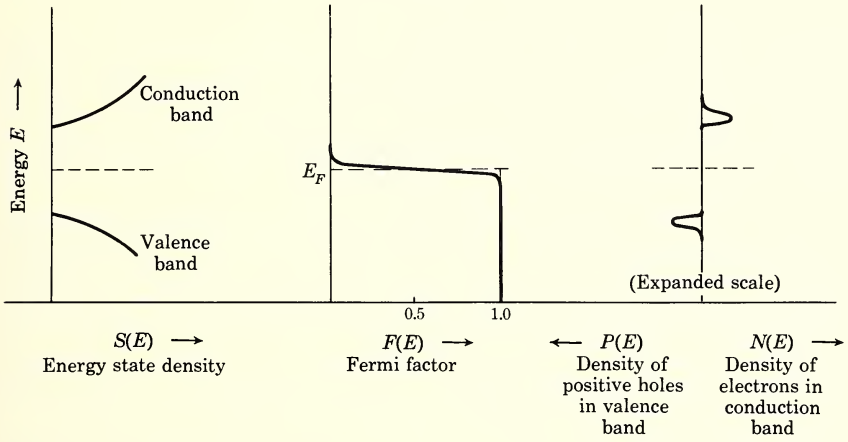


FIG. 2–5. Graphical illustration of the number of electrons and holes in an intrinsic semiconductor showing the position of the Fermi level.

available number of states and the probability of their occupancy. Equation (2–11) is repeated here for convenience.

$$N(E)\,dE = S(E)F(E)\,dE \tag{2–11}$$

This relationship is illustrated in Fig. 2–5. The left-hand curve is a plot of Eqs. (2–8) and (2–9) and shows the density of available electron states close to the forbidden gap for both the valence and conduction bands. The central curve shows the Fermi factor at room temperature, and the right-hand curve is obtained by taking the product of the values shown in the two preceding curves at all energy levels. The density of electrons in the conduction band as given by Eq. (2–11) is shown to the right of the energy axis. The density of

positive holes in the valence band is shown to the left of the energy axis. This number can be represented by $P(E)\,dE$ in an equation which uses $1 - F(E)$ in place of $F(E)$ in Eq. (2–11), i.e.,

$$P(E)\,dE = F(E)[1 - F(E)]\,dE \tag{2–13}$$

Under certain conditions an approximation to the Fermi factor can be made. For a room temperature of 300° K, for example, the value of kT is 25.8 millielectron volts. At an energy level E_1 which is more than 100 millielectron volts (say) above the Fermi level,

$$\exp\left(\frac{E_1 - E_F}{kT}\right) \gg 1$$

and Eq. (2–10) can be written as

$$F(E_1) \approx \exp\left(\frac{E_F - E_1}{kT}\right) \tag{2–14}$$

The right-hand side of Eq. (2–14) is known as the *Boltzmann factor* because it corresponds with the relationship obtained from classical Maxwell-Boltzmann statistics, which apply to a perfect gas. We see from Eq. (2–14) that the probability of occupancy drops exponentially with energy at levels well above E_F.

The probability of finding an electron state unoccupied in the valence band has already been given as the difference between unity and the Fermi factor at that energy. Assuming E_2 is an energy level in the valence band which is far removed from the Fermi level,

$$1 - F(E_2) = 1 - \frac{1}{1 + \exp\left[(E_2 - E_F)/kT\right]} \tag{2–15}$$

where $E_F - E_2 \gg kT$. In Eq. (2–15), the exponential term is much less than unity, and so by the binomial expansion,

$$1 - F(E_2) \approx 1 - \left[1 - \exp\left(\frac{E_2 - E_F}{kT}\right)\right]$$

$$= \exp\left(\frac{E_2 - E_F}{kT}\right) \tag{2–16}$$

Equation (2–16) indicates that as E_2 becomes more negative the probability of finding holes decreases exponentially.

This information allows us to calculate the position of the Fermi level in an intrinsic semiconductor. First, it should be noted that the Fermi level will be somewhere in the forbidden gap since nearly all the valence band states are occupied, and only very few electrons will be found in the conduction band at room temperatures. Next, Eqs. (2–8) and (2–9) give the required information of the energy state densities in the two bands, and finally, Eqs.

(2–14) and (2–16) give the approximate probability factors to be used in the calculation.

Consider energy levels E_1 in the conduction band and E_2 in the valence band symmetrically placed about the center of the gap. For an energy range dE in both cases, the number of electrons in the conduction band is

$$N(E_1)\,dE = S(E_1)F(E_1)\,dE \tag{2–17}$$

and the number of positive holes in energy range dE in the valence band is

$$P(E_2)\,dE = S(E_2)[1 - F(E_2)]\,dE \tag{2–18}$$

Assuming for simplicity that $m_c = m_v$ in Eqs. (2–8) and (2–9),

$$S(E_1) = S(E_2) \tag{2–19}$$

Thus,

$$\frac{N(E_1)}{P(E_2)} = \frac{F(E_1)}{1 - F(E_2)} = \frac{\exp\,[(E_F - E_1)/kT]}{\exp\,[(E_2 - E_F)/kT]} \tag{2–20}$$

assuming $E_1 - E_F \gg kT$ and $E_F - E_2 \gg kT$.

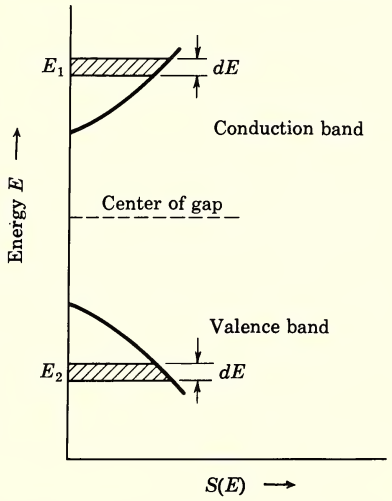


FIG. 2–6. Curve of $S(E)$ versus E for an intrinsic semiconductor showing the location of elements spaced symmetrically about the center of the forbidden gap.

Equation (2–20) applies to all symmetrically placed elements on Fig. 2–6. Hence, the ratio of the numbers of electrons and positive holes for any element is the same as the ratio of the total numbers of electrons and positive holes in the semiconductor, which is unity for intrinsic material. Thus,

$$\frac{N(E_1)}{P(E_2)} = \frac{n_i}{p_i} = 1 \tag{2–21}$$

Therefore, in Eq. (2–20)

$$\frac{E_F - E_1}{kT} = \frac{E_2 - E_F}{kT}$$

or

$$E_F = \frac{E_2 + E_1}{2} \tag{2–22}$$

It will be recalled that the energy levels E_1 and E_2 were taken symmetrically about the center of the forbidden gap. *Equation* (2–22) *shows that the Fermi level lies at the center of the forbidden gap for intrinsic material and that it is independent of temperature.* The error introduced by putting $m_c = m_v$ is small in practice (see Problem 2–7).

2–4 *n*-type Semiconductor

When atoms of group 5 material are introduced into an otherwise pure crystal of germanium or silicon, additional energy levels are introduced just below the bottom of the conduction band in the forbidden gap. The curve of available energy states versus energy is no longer symmetrical about the center, and the Fermi level is no longer at the center of the gap.

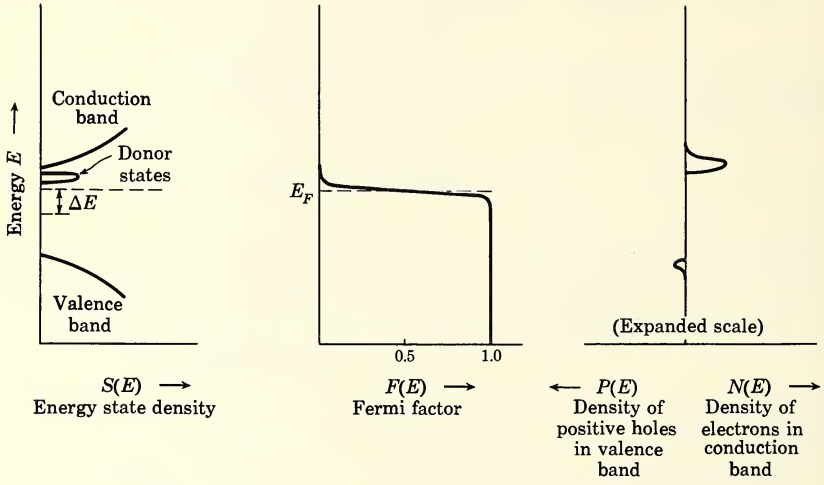


FIG. 2–7. The position of the Fermi level in an *n*-type semiconductor.

In Fig. 2–7, the Fermi level is shown at an energy ΔE above the center of the gap. By comparison with the case of the intrinsic semiconductor where the Fermi level is located at the center of the gap, the Fermi factor is increased by the ratio $\exp(\Delta E/kT)$ everywhere in the conduction band [see Eq. (2–14)]. Thus, the total number of electrons in the band (which was n_i for the intrinsic case) is now given by

$$n = n_i \exp\left(\frac{\Delta E}{kT}\right) \tag{2–23}$$

For an *n*-type semiconductor, $n > n_i$, and so ΔE is positive, i.e., the Fermi level for an *n*-type semiconductor lies above the center of the gap.

The energy level of the donor states lies just below the bottom of the conduction band, and the Fermi factor at this energy is low. The donor states will, therefore, be almost empty, and the donor electrons will be found in the conduction band. Note that the probability of occupancy in the conduction band is even lower than in the donor states, but there are many more states present in the conduction band.

When the density of donor atoms in the material is high, the conductivity is almost entirely due to donor electrons in the conduction band as explained in Secs. 1–19 and 2–8. Under these conditions,

$$n = N_d = n_i \exp\left(\frac{\Delta E}{kT}\right) \tag{2–24}$$

and so the position of the Fermi level can be calculated if N_d and n_i are known.

The number of holes in the valence band is small in *n*-type material. This is shown diagrammatically in Fig. 2–7 where it is seen that raising the Fermi level above the center of the forbidden gap decreases the value of $1 - F(E)$ at the top of the valence band.

2–5 *p*-type Semiconductor

When group 3 atoms are introduced into a germanium or silicon crystal, the acceptor energy level is just above the valence band in the forbidden gap. The number of positive holes in the valence band is greater than the number of electrons in the conduction band, and the Fermi level is below the center of the forbidden gap as shown in Fig. 2–8.

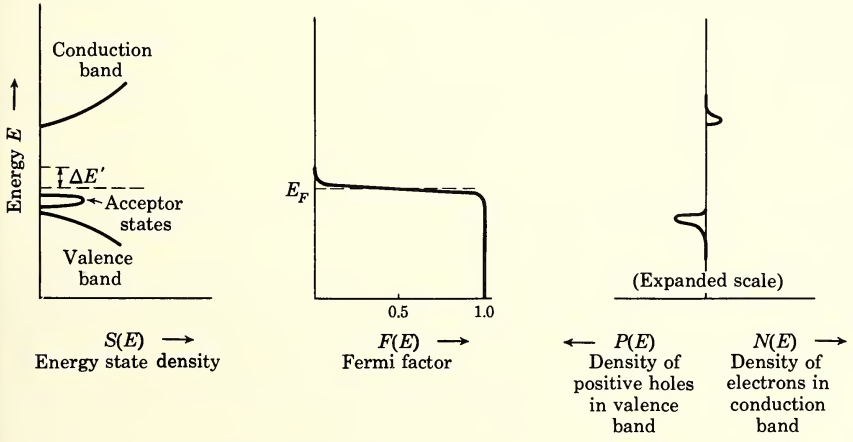


FIG. 2–8. The position of the Fermi level in a *p*-type semiconductor.

If $\Delta E'$ is the energy by which the Fermi level is displaced down from the center of the gap in this material, the total number of holes in the valence band can be written as

$$p = n_i \exp\left(\frac{\Delta E'}{kT}\right) \tag{2–25}$$

by comparison with the arguments of the previous section. For material

which has so many added acceptor atoms that the conductivity is mainly due to holes, $p = N_a$, and

$$N_a = n_i \exp\left(\frac{\Delta E'}{kT}\right) \tag{2-26}$$

The number of electrons in the conduction band is small, as shown in Fig. 2–8.

2–6 Variation of the Fermi level with Temperature

For intrinsic material, raising the temperature increases the number of both electrons and positive holes since more energy is available to allow a valence electron to break its bond. The number of electrons is still equal to the number of positive holes, however, and so the Fermi level remains approximately at the center of the forbidden energy gap whatever the temperature.

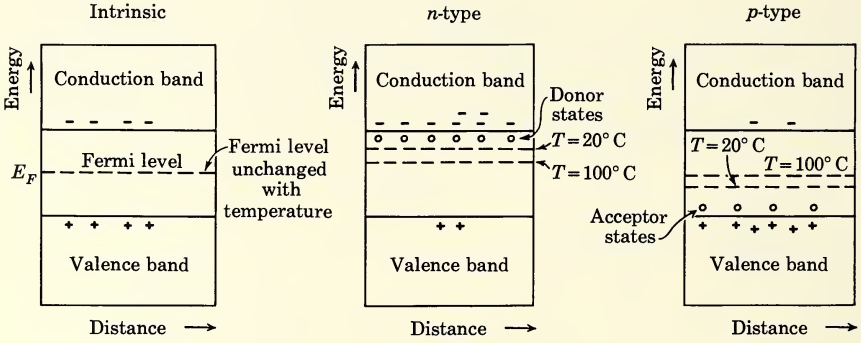


FIG. 2–9. Effect of temperature on Fermi level for typical intrinsic n- and p-type semiconductors.

In an n-type crystal, there are electrons in the conduction band that have come from two different sources. Some of the electrons will be those from the group 5 impurity atoms, and these are easily separated from their parent atom. Their number does not vary much as the temperature is altered within the range of temperatures where semiconducting devices are commonly used (0° to 50° C). The other electrons in the conduction band are present because of the breaking of a covalent bond. Such "intrinsic effect" electrons will increase in number as the temperature is raised, and so their proportion of the total number will increase. Thus as the temperature rises, the material becomes more intrinsic and the Fermi level moves closer to the intrinsic position, that is, to the center of the forbidden gap. Hence as the temperature is raised, the material becomes increasingly nearer to intrinsic material in its electrical properties.

For p-type material, the Fermi level at room temperature is below the center of the forbidden gap. As the temperature is increased, this material

also becomes increasingly intrinsic for the same reason, and the Fermi level rises until it approaches the center of the gap. Thus both *p*- and *n*-type material become more like intrinsic material at high temperatures. *This places a limit on the operating temperature of a semiconducting device.*

2–7 The Law of Mass Action

The number of electrons in an energy range dE in the conduction band at a level E has previously been given as

$$N(E)\,dE = S(E)\exp\left(\frac{E_F - E}{kT}\right)dE \tag{2–27}$$

where $E - E_F \gg kT$. The total number of electrons in the conduction band is

$$n = \int_{E_c}^{\infty} N(E)\,dE = \int_{E_c}^{\infty} S(E)\exp\left(\frac{E_F - E}{kT}\right)dE \tag{2–28}$$

The upper limit of the integral is put as infinity for convenience in the integration. This limit will certainly include all the electrons in the conduction band. Now the value of $S(E)$ has been given previously as

$$S(E) = K_c(E - E_c)^{1/2} \tag{2–8}$$

for the bottom of the conduction band only, where

$$K_c = 2^{7/2}m^{3/2}\pi h^{-3}$$

The exponential term of Eq. (2–28) dominates the equation for energies more than a few kT above the bottom of the conduction band, and the value of the integrand soon drops to zero upon moving upwards through the band. Hence, all the electrons will be found at the bottom of the band, and so the limit can be used even when Eq. (2–8) is substituted in the integral. Thus,

$$\begin{aligned} n &= K_c\int_{E_c}^{\infty}(E - E_c)^{1/2}\exp\left(\frac{E_F - E}{kT}\right)dE \\ &= K_c\exp\left(\frac{E_F - E_c}{kT}\right)\int_{E_c}^{\infty}(E - E_c)^{1/2}\exp\left(\frac{E_c - E}{kT}\right)dE \end{aligned} \tag{2–29}$$

This integral may be put in the form

$$\int_0^{\infty} x^{1/2}e^{-x}\,dx = \frac{\pi^{1/2}}{2}$$

and so

$$n = 2^{-1}K_c\pi^{1/2}k^{3/2}T^{3/2}\exp\left(\frac{E_F - E_c}{kT}\right) \tag{2–30}$$

Similarly for holes in the valence band it can be shown that

$$p = 2^{-1}K_v\pi^{1/2}k^{3/2}T^{3/2} \exp\left(\frac{E_v - E_F}{kT}\right) \tag{2-31}$$

where

$$K_v = 2^{7/2}m^{3/2}\pi h^{-3}$$

Multiplying Eq. (2–30) by Eq. (2–31),

$$np = K_1^2T^3 \exp\left(\frac{E_v - E_c}{kT}\right)$$

$$= K_1^2T^3 \exp\left(\frac{-E_g}{kT}\right) \tag{2-32}$$

where $E_g = E_c - E_v$, and is the *width of the forbidden gap*, and

$$K_1^2 = \frac{K_cK_v\pi k^3}{4} \tag{2-33}$$

From Eq. (2–32), *the product np is seen to be independent of the position of the Fermi level and is thus the same value for p-type, n-type or intrinsic material.* Using the subscript *i* for intrinsic material,

$$np = n_ip_i = n_i^2 \tag{2-34}$$

Equation (2–34) is known as the *law of mass action.* From Eqs. (2–32) and (2–34),

$$n_i = K_1T^{3/2} \exp\left(\frac{-E_g}{2kT}\right) \tag{2-35}$$

which allows us to calculate n_i for a given material if K_1 and E_g are known. In the previous sections of this chapter, we put $m_c = m_v$ without incurring serious error. In this calculation, however, the product of m_c and m_v is involved in the constant K_1 of Eq. (2–35), and m_c cannot be equated to m_v. If the energy gap and the intrinsic density of the carriers are known by experiment, the product of the effective masses can be calculated from Eq. (2–35) (see Problem 2–8).

For germanium at 300° K, $E_g = 0.72$ ev, and $n_i = 2.5 \times 10^{13}$ cm^{-3}. For silicon at 300° K, $E_g = 1.1$ ev, and $n_i = 1.6 \times 10^{10}$ cm^{-3}.

2–8 Mobile and Immobile Charges

Consider a piece of semiconducting material containing atoms of both group 3 and group 5. Let N_a and N_d be the density of atoms of the two groups respectively. The material will be *p*-type or *n*-type according to which kind of impurity atom predominates. Let the density of positive holes in the material be p and the density of electrons be n.

In germanium or silicon at room temperature it is found that the energy state of the fourth (missing) valence electron of the group 3 (acceptor) atom is sufficiently close to the valence band that every acceptor ion has an extra electron associated with it (see Sec. 1–19 and Fig. 1–20). Thus there is a resultant negative charge $(-e)$ associated with each acceptor atom in the crystal lattice at room temperature. Similarly, every group 5 donor atom donates its extra electron into the conduction band and a positive charge $(+e)$ is associated with every donor atom (see Fig. 1–18).

The crystal as a whole must be electrically neutral since it was formed from neutral group 3, 4, or 5 atoms. Therefore, a charge neutrality equation may be written summing the negative charges as $ne + N_a e$ and the positive charges as $pe + N_d e$. Hence,

$$n + N_a = p + N_d \tag{2–36}$$

From Eq. (2–34),

$$n = \frac{n_i^2}{p}$$

Hence,
$$\frac{n_i^2}{p} + N_a = p + N_d$$

or
$$p^2 + (N_d - N_a)p - n_i^2 = 0 \tag{2–37}$$

The solution of Eqs. (2–34) and (2–36) is, therefore,

$$p = \frac{-(N_d - N_a) \pm \sqrt{(N_d - N_a)^2 + 4n_i^2}}{2}$$

$$n = \frac{(N_d - N_a) \pm \sqrt{(N_d - N_a)^2 + 4n_i^2}}{2} \tag{2–38}$$

Assume that the material is n-type and that $N_d - N_a \gg 2n_i$. Then in Eq. (2–38) the positive sign in front of the square root is selected since n must be positive. Hence,

$$n \approx N_d - N_a$$

and from Eq. (2–34),

$$p \approx \frac{n_i^2}{N_d - N_a} \tag{2–39}$$

When the density of acceptor atoms is higher than donor atoms in the crystal, and assuming that $N_a - N_d \gg 2n_i$, Eq. (2–38) becomes

$$p = N_a - N_d$$

and
$$n = \frac{n_i^2}{N_a - N_d} \tag{2–40}$$

Equations (2–39) and (2–40) show that it is the excess of one type of impurity density over the other that determines the type of material. When

$N_a = N_d$, the material becomes intrinsic again provided the donor and acceptor atoms are not present in such numbers as to destroy the lattice properties of the group 4 crystal.

For n-type material where no acceptor atoms are present and $N_d \gg 2n_i$,

$$n = N_d$$

$$p = \frac{n_i^2}{N_d} \tag{2–41}$$

Since N_d is relatively high, the contribution of the minority carriers to the conductivity can be ignored, and

$$\sigma = ne\mu_n + pe\mu_p \approx N_d e\mu_n \tag{2–42}$$

For strongly p-type material where no donor atoms are present, the corresponding equations are

$$p = N_a$$

$$n = \frac{n_i^2}{N_a} \tag{2–43}$$

and

$$\sigma = ne\mu_n + pe\mu_p \approx N_a e\mu_p \tag{2–44}$$

Example:

Germanium has a donor-type impurity added to the extent of one atom per 10^8 germanium atoms. What effect does this have on the conductivity of the material at 300° K?

For intrinsic germanium at 300° K from Table 1–2 and Figs. 1–21 and 1–22,

$$n_i^2 = 6.25 \times 10^{26}\ \text{cm}^{-3}$$
$$\mu_n = 3900\ \text{cm}^2\ \text{volt}^{-1}\ \text{sec}^{-1}$$
$$\mu_p = 1900\ \text{cm}^2\ \text{volt}^{-1}\ \text{sec}^{-1}$$

The conductivity of intrinsic germanium is

$$\sigma_i = n_i e(\mu_n + \mu_p)$$
$$= 2.5 \times 10^{13}(1.6 \times 10^{-19})5800$$
$$= 0.0232\ \text{ohm}^{-1}\ \text{cm}^{-1}$$

There are 4.42×10^{22} atoms of germanium per cubic centimeter (see Appendix), and so the number of donor atoms per cubic centimeter is 4.42×10^{14}. N_d is, therefore, about 18 times n_i, and we may use the simplified form of Eq. (2–38) without introducing an error of more than 1 per cent. This is not significant, since semiconductor properties can seldom be measured with an

accuracy of better than 3 per cent. Thus,

$$n = N_d = 4.42 \times 10^{14}\ \text{cm}^{-3}$$

$$p = \frac{n_i^2}{N_d} = \frac{6.25 \times 10^{26}}{4.42 \times 10^{14}} = 1.41 \times 10^{12}\ \text{cm}^{-3}$$

and Eq. (2–42) becomes

$$\sigma \approx N_d e \mu_n = 4.42 \times 10^{14}(1.6 \times 10^{-19})3800$$

$$= 0.269\ \text{ohm}^{-1}\,\text{cm}^{-1}$$

where the mobility is obtained from Fig. 1–21.

This example illustrates several important points. First, the addition of only 1 part of donor impurity to 10^8 parts of germanium increases the conductivity of a crystal approximately ten-fold, showing how sensitive the conductivity is to donor (or acceptor) type impurity. Secondly, although this type of calculation aids us in understanding the mechanism of semiconductors, it is difficult to carry out for high impurity densities where the mobility is a function of the conductivity. Finally, the experimental results given in Figs. 1–21 through 1–24 provide a ready-made solution to this type of problem.

2–9 The Hall Effect

It is often necessary to determine whether a material is *n*-type or *p*-type. Measurement of the conductivity of a specimen will not give this information

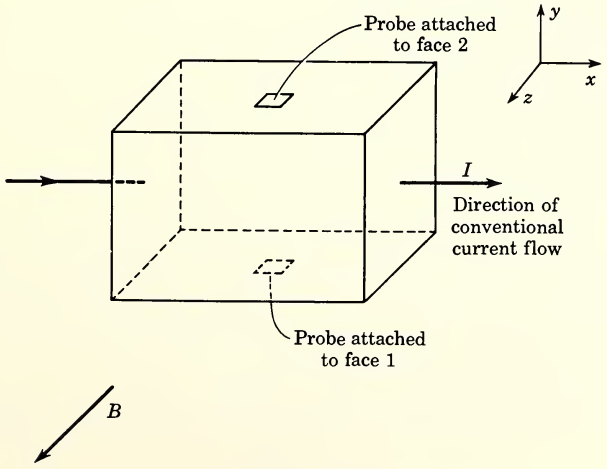


FIG. 2–10. Illustration of the Hall effect.

since it cannot distinguish between positive hole and electron conduction. The *Hall effect* can be utilized to distinguish between the two types of carrier, and it also allows the density of the charge carriers to be determined. In conjunction with a measurement of conductivity the mobility of the carriers can be found.

In 1897, Hall discovered that when a current was passed through a slab of material in the presence of a transverse magnetic field, a small potential difference was established in a direction perpendicular to both the current flow and the magnetic field. Figure 2–10 shows the current flowing in the positive x direction with the flux density, B, in the positive z direction and the potential difference appearing in the y direction. Probes may be attached to faces 1 and 2 and the potential difference may be measured by a high sensitivity d-c vacuum tube voltmeter of very high input resistance.

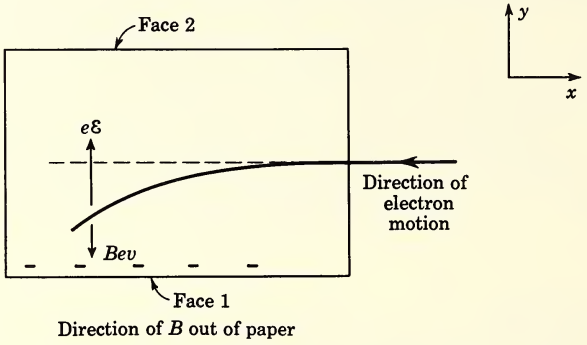


FIG. 2–11. Motion of electrons in an n-type semiconductor (Hall effect).

The Hall effect may be explained by reference to Fig. 2–11 which shows the front face of the slab only. Assuming that the material is an n-type semiconductor, the current flow consists almost entirely of electrons moving from right to left. This corresponds to the direction of conventional current from left to right as in Fig. 2–10.

If v is the velocity of electrons at right angles to the magnetic field, there is a downward force on each electron of magnitude Bev. This causes the electron current to be deflected in a downwards direction and causes a negative charge to accumulate on the bottom face of the slab (face 1). A potential difference is therefore established from top to bottom of the specimen with the bottom face negative. This potential difference causes a field $\mathscr{E}$ in the negative y direction, and so there is a force of $e\mathscr{E}$ acting in the upward direction on the electron. Equilibrium occurs when

$$e\mathscr{E} = Bev$$

or

$$\mathscr{E} = Bv \tag{2–45}$$

The Hall coefficient is defined to be

$$R_n = -\frac{\mathscr{E}}{BJ} \tag{2–46}$$

where J is the current density of the electron stream, and the negative sign is used because the electric field is in the negative y direction. Now $J = nev$, thus

$$R_n = -\frac{\mathscr{E}}{BJ} = -\frac{v}{nev} = -\frac{1}{ne} \tag{2–47}$$

All three quantities on the right-hand side of Eq. (2–46) can be measured, and so the Hall coefficient and the carrier density n can be found.

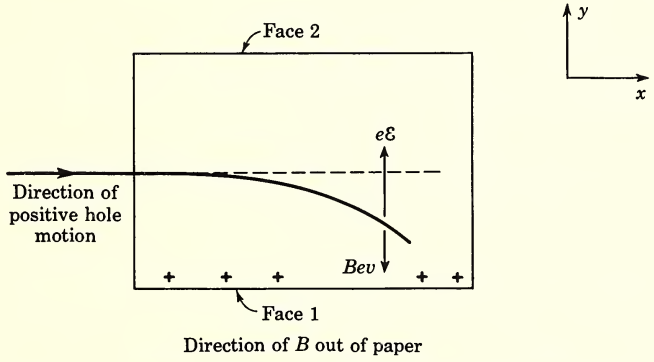


FIG. 2–12. Motion of positive holes in a p-type semiconductor (Hall effect).

Figure 2–12 shows the conditions that exist in a similar p-type specimen when the current is carried entirely by positive holes. If the conventional current flow is in the same direction, the deflection of the holes is downwards, *but now the carriers have a positive charge.* Thus the bottom face becomes positive, and the potential difference and the field are in the opposite direction from the previous case. The sign of the charge and the direction of the field are both changed so Eq. (2–45) still applies, i.e.,

$$\mathscr{E} = Bv \tag{2–45}$$

The Hall coefficient for hole conduction is now positive and is given by

$$R_p = \frac{\mathscr{E}}{BJ} \tag{2–48}$$

Substituting $J = pev$ where p is the positive hole density gives

$$R_p = \frac{1}{pe} \tag{2–49}$$

Thus the sign of the Hall voltage indicates which carriers predominate in the material, and the magnitude of the potential difference may be used to calculate the density of charge carriers.

The analysis given above applies only when the charge carriers are free of attractive forces in the energy bands and when they move with a steady drift velocity v. This is not the case in a semiconductor, and a computation of the average speed leads to the conclusion that

$$R_n = \frac{-3\pi}{8ne} = \frac{-1.18}{ne} \tag{2-50}$$

$$R_p = \frac{3\pi}{8pe} = \frac{1.18}{pe} \tag{2-51}$$

For n-type material, the conductivity is given by

$$\sigma_n = ne\mu_n$$

and in p-type material

$$\sigma_p = pe\mu_p$$

Hence,
$$\mu_n = \frac{\sigma_n}{ne} = \frac{-\sigma_n R_n}{1.18} \tag{2-52}$$

and
$$\mu_p = \frac{\sigma_p}{pe} = \frac{\sigma_p R_p}{1.18} \tag{2-53}$$

Therefore, the mobility of the charge carriers may be found if the conductivity and the Hall coefficient are known.

BIBLIOGRAPHY

General

Dekker, Adrianus J., *Solid State Physics*, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1957

Middlebrook, R. D., *An Introduction to Junction Transistor Theory*, New York: John Wiley & Sons, Inc., 1957

Spenke, Eberhard, *Electronic Semiconductors*, New York: McGraw-Hill Book Company, Inc., 1958

Sproull, Robert L., *Modern Physics*, New York: John Wiley & Sons, Inc., 1956

Hall effect

Dunlap, W. C., *An Introduction to Semiconductors*, New York: John Wiley & Sons, Inc., 1957

Kittel, Charles, *Introduction to Solid State Physics*, 2d ed., New York: John Wiley & Sons, Inc., 1956

Shive, J. N., *Properties, Physics and Design of Semiconductor Devices*, New York: D. Van Nostrand Company, Inc., 1959

PROBLEMS

2–1 A metal has its Fermi level 8.95 ev above the bottom of the conduction band. Find the momentum of an electron at the Fermi level assuming the free mass of an electron applies. How many energy states per unit volume are present below the Fermi level?

2–2 What is the range in energy in multiples of kT over which the Fermi factor varies from 0.99 to 0.01? What is this energy range in electron volts at 300° K?

2–3 What fundamental concepts are used in the derivation of the energy density of states in the conduction band of a metal as a function of energy?

2–4 Determine the average energy of an electron in the conduction band of a metal at 0° K as a function of the Fermi level.

ans. (3/5) E_F

2–5 What is the percentage error introduced by making the approximation in Eq. (2–14) that $E_1 - E_F \gg kT$ (in particular, that 100 millielectron volts $\gg$ 25.8 millielectron volts)?

2–6 What are the significant differences between crystalline solid state materials and gases that necessitate the use of different statistical treatments (Fermi-Dirac for solids and Maxwell-Boltzmann for gases) in the evaluation of electrical phenomena?

2–7 Recalculate Eq. (2–22) using m_c and m_v for the mass of an electron in the conduction band and a hole in the valence band respectively. If $m_c/m_v = 1.5$, what is the displacement of the Fermi level from the center of the forbidden gap when $T = 300°$ K?

2–8 At 300° K, the width of the forbidden energy gap in germanium is 0.72 ev and the intrinsic carrier density is 2.5×10^{13} cm^{-3}. Calculate the value $m_c m_v/m^2$, where m is the free mass of an electron.

2–9 Carefully distinguish among intrinsic, n-type and p-type semiconductor. Assuming $m_c = m_v$, what is the position of the Fermi level at 300° K (a) in intrinsic germanium (b) in n-type germanium where $N_d = 10^{15}$ cm^{-3} (c) in p-type germanium where $N_a = 10^{13}$ cm^{-3}?

2–10 Calculate the resistivity of intrinsic silicon at 300° K from the data of Table 1–2.

2–11 Calculate the density of donor atoms which has to be added to intrinsic germanium to produce *n*-type material of conductivity 0.19 ohm cm. It is given that the mobility of electrons in the *n*-type semiconductor is 3250 cm^2 $volt^{-1}$ cm^{-1}.

ans. 1.01×10^{16} cm^{-3}

2–12 Explain how the Hall effect shows whether holes or electrons predominate in a semiconductor. Suggest an experimental arrangement for measuring Hall voltage giving values of current, magnetic field, and voltage you would expect to find for a small specimen of *n*-type germanium of resistivity 5 ohm cm.

2–13 An *n*-type germanium specimen has a donor density of 10^{15} cm^{-3}. It is arranged in a Hall effect experiment where $B = 0.5$ webers m^{-2} and $J = 500$ amps m^{-2}. What is the Hall voltage if the specimen is 3 mm thick?

ans. 5.5 millivolts

3

Junctions between Materials

Before discussing semiconductor junctions, we will briefly consider the less complex case of a junction between two metals. It is shown that there is an initial passage of charge across the junction which causes a change in potential between the two sides. A semiconductor junction is then examined in more detail, and the contact potential, the junction capacitance, and the current versus voltage characteristic are derived.

3–1 Work Function in a Metal

Figure 3–1 shows the energy state density curve and the position of the Fermi level for a typical metal. The curve of $S(E)$ versus energy is of the form

$$S(E) = K(E)^{1/2} \tag{3–1}$$

where the energy, E, is measured from the bottom of the band, and $K = 2^{7/2} m^{3/2} \pi h^{-3}$.

The Fermi factor, $F(E)$, has the value given in Eq. (2–10) which is repeated here for convenience

$$F(E) = \frac{1}{1 + \exp\left[(E - E_F)/kT\right]}$$

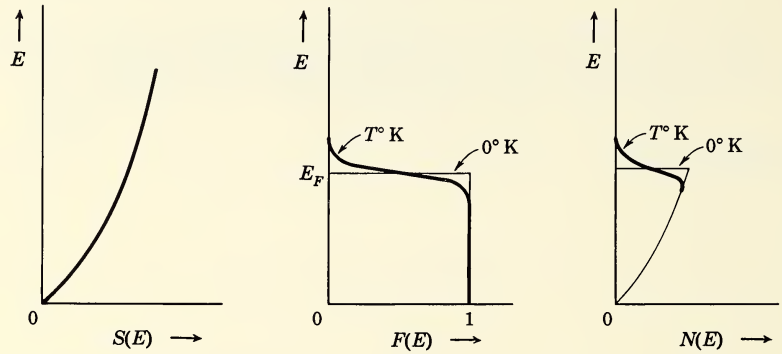


FIG. 3–1. Position of the Fermi level for a metal.

The curve of the number of occupied states, $N(E)$, versus energy which is on the right-hand side of Fig. 3–1 is obtained from Eq. (2–11), i.e.,

$$N(E)\,dE = S(E)F(E)\,dE$$

For a metal, the total number of conduction electrons remains constant as the temperature is changed. The variation of $F(E)$ with temperature has been discussed in Chap. 2. However, it is found that, although the shape of the curve of Fermi factor plotted against energy is dependent upon temperature, the curve alters in such a way that the Fermi level is substantially constant. The Fermi level of a metal, therefore, is a useful reference level.

In any metal the energy states at the top of the band are not completely full. Because there is a continuous band of energy states, electrons are free to move about within the metal, and the electrical conductivity is high. This leads to a picture of the potential energy variation in a metal which ignores the potential energy variations close to an atom but has the merit of simplicity. Sommerfeld first introduced this simple model of electrons in a metal as

given in Fig. 3–2. Here electron levels up to the Fermi level E_F are shown filled. The *vacuum level* in the diagram refers to the energy of an electron at rest outside the metal.

The vacuum level, E_S, must be higher than the Fermi level, E_F, for a metal at room temperatures, otherwise electrons would be thermionically emitted (see below). The difference between E_S and E_F is known as the *work function*, ϕ, of the material. It is the energy required to raise one electron from the Fermi level inside the metal to the vacuum level outside. The value of the work function is usually between 2 and 5 ev.

The Fermi level is measured from the bottom of the energy band and can be calculated for a particular metal if the valence, density, and atomic weight of the metal are known. The valence of a metal is the number of electrons per atom in the outer energy band, and so the product of the valence and the number of atoms per unit volume gives the total number of electrons per unit

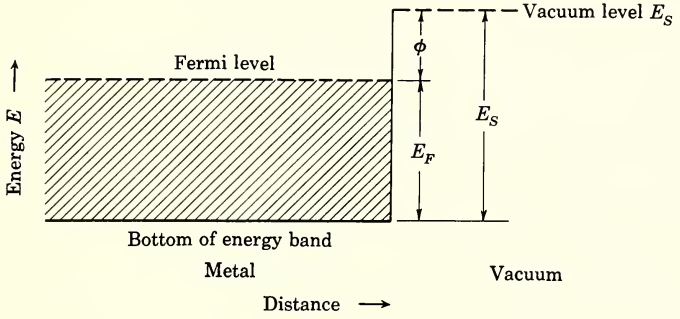


FIG. 3–2. Simple picture of electrons in a metal. Conditions close to the surface of the metal are not shown.

volume which are occupying energy states in this band. If this number is N, from Eqs. (3–1) and (2–11),

$$N = \int_0^{\infty} KE^{1/2}F(E)\,dE = \int_0^{E_F} KE^{1/2}\,dE \tag{3–2}$$

at $T = 0°$ K, since $F(E) = 1$ up to E_F and all states will be filled up to the Fermi level and unfilled above that level. Hence,

$$N = \frac{2KE_F^{3/2}}{3}$$

or

$$E_F = \left(\frac{3N}{2K}\right)^{2/3} \tag{3–3}$$

Table 3–1 gives the values of the Fermi level and the work function for several metals.

TABLE 3–1. Valence, Fermi level, and work function for five common metals.

Metal	*Valence*	E_F	ϕ
Al	3	11.7 ev	4.20 ev
Ag	1	5.5	4.46
Cu	1	7.0	4.45
K	1	2.1	2.22
Na	1	3.1	2.28

As the temperature of a metal (or semiconductor) is raised, it is found that electrons will be emitted from the surface. This phenomenon is known as *thermionic emission* and is of great importance in thermionic vacuum tubes where it supplies the electron current in the device. The electron state distribution curve of Fig. 3–3 shows that at high temperatures there is an appreciable number of electrons in the "tail" of the curve above the vacuum level, E_S. This number can be calculated at any given temperature from Eq. (2–11). These electrons have sufficient energy to escape from the metal surface over the potential barrier ϕ (see Fig. 3–2). However, all electrons having an energy greater than E_S will not necessarily be emitted from the surface. The calculation of the emitted current takes into account only those electrons which are traveling with a velocity normal to the surface sufficient to overcome the energy barrier of the work function, ϕ.

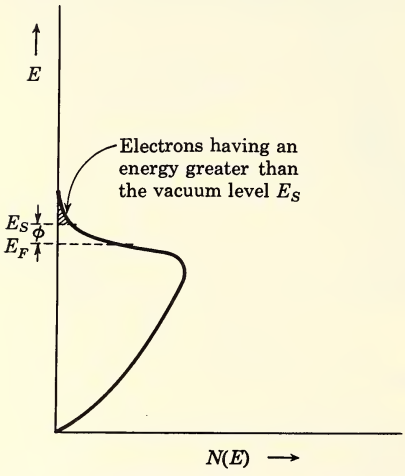


FIG. 3–3. Diagram showing electrons in the tail of the distribution curve which have an energy greater than E_S.

This example briefly indicates how the Fermi-Dirac statistics can be used to explain a well-known phenomenon in physical electronics. In the remaining sections of this chapter, the Fermi level will be shown to be an important tool in the analysis of current flow across metals and semiconductor junctions.

3–2 Junctions between Metals

In Fig. 3–4 is shown the energy band diagram at 0° K of two different metals having work functions ϕ_1 and ϕ_2. The shaded areas represent the filled

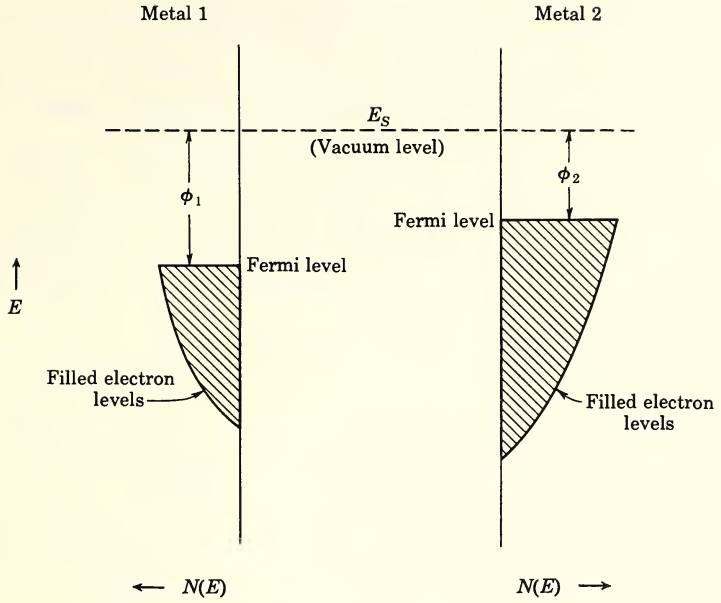


FIG. 3–4. Energy band diagram of two metals with work functions ϕ_1 and ϕ_2 (metals are not in contact).

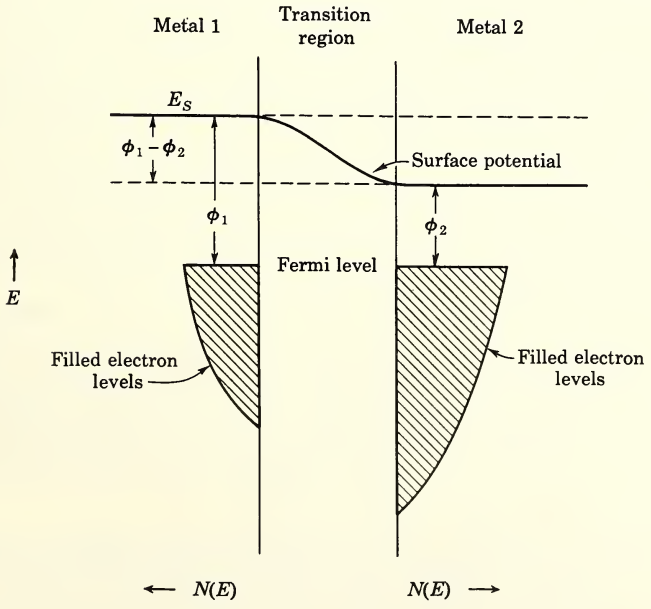


FIG. 3–5. Junctions of two metals at 0° K showing the contact potential (metal 1 at ground potential).

electron levels in each metal, and the metals are assumed to be separated. The reference level E_S is common.

Figure 3–5 shows the energy band diagram when the two metals have a common junction. The junction is assumed to be clean and regular so that the crystalline properties of the two materials are altered as little as possible. Since the electrons are virtually free in both metals, there is a transfer of electrons from metal 2 to metal 1 until the Fermi levels coincide. In this process, metal 1 acquires electrons and becomes negative, and metal 2 loses electrons and becomes positive. The difference in energy possessed by an electron on the two sides of the junction is $\phi_1 - \phi_2 = eV_B$. Thus V_B is the potential difference between the two metals which occurs because of the transfer of charge across the junction. V_B is known as the *contact potential* between the metals.

For metals at 0° K it can easily be seen that, since the highest filled electron states coincide with the Fermi level, the two Fermi levels on either side of the junction will line up. At any other temperature, there are some filled electron states above the Fermi levels, but it may be shown thermodynamically that, at any temperature, the Fermi levels of any two materials will coincide when the two materials have a common junction.

3–3 Semiconductor Junctions with No Applied Voltage

Semiconductor junctions are commonly formed between *p*- and *n*-type material. It is assumed that the group 4 crystalline structure is unaltered across the junction and that only the type of impurity changes. Thus, junctions having the properties described here cannot be made by placing two pieces of semiconducting material together, because surface films and other irregularities would produce major discontinuities in the crystal structure. Junctions can be made by diffusing the required impurity density into the crystal, by an alloying process, or by growing a crystal with an impurity density which is a function of distance. These three methods of construction are briefly discussed in Sec. 5–6.

Junctions may be *abrupt* or *graded*. An abrupt junction, which is also known as a *step junction*, is of the type shown later in Fig. 3–7 where the acceptor or donor density is a constant in the material up to the boundary. In a graded junction, the impurity density varies with distance away from the boundary in some manner. Both types of junction are used extensively in the manufacture of diodes and transistors. Only the abrupt type of junction will be discussed in detail in this chapter.

In a metal, the electrons which move across the boundary to bring the Fermi levels in line are drawn from a very thin layer close to the junction since there is a high density of free electrons in the metal (about 10^{22} electrons per cubic centimeter). For a semiconductor, however, there may be only about 10^{15} free electrons per cubic centimeter, in a typical case, and so there

is a much wider region which is affected by the flow of charge. Electrons in the conduction band on the *n*-type side, which were supplied from the group 5 impurity atoms, travel across the junction and leave the positively ionized group 5 atoms unneutralized (Fig. 3–6). Consequently there is a *positively*

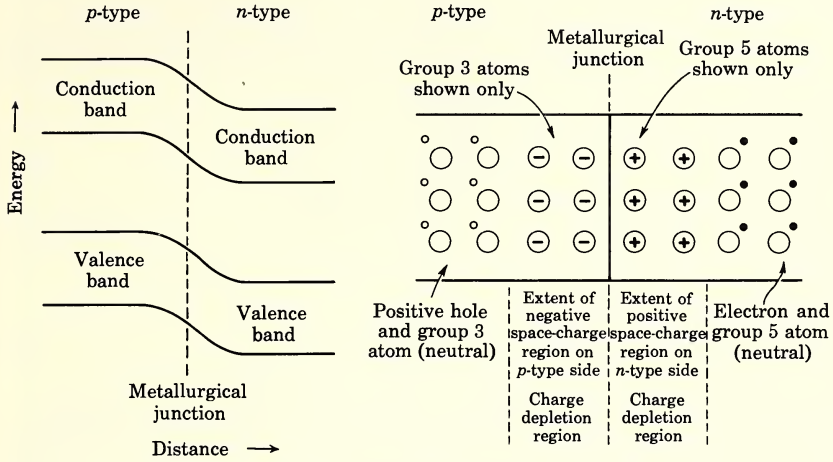


FIG. 3–6. The *p–n* junction.

charged region adjacent to the junction in the *n*-type material. On the *p*-type side, the electrons which have traversed the boundary recombine with the positive holes in the valence band, which predominate because of the presence of group 3 atoms. Close to the junction on the *p*-type side, there is a layer of unneutralized negatively ionized group 3 ions which form a *negatively charged region*. Both the group 5 positive ions (in the *n*-type material) and the group 3 negative ions (in the *p*-type material) are immobile since they are bound in the crystal lattice. Close to the metallurgical junction, therefore, there are very few mobile charge carriers, and this location is referred to as the *charge depletion region*. In fact, it is a charged or *space charge region* because of the presence of the immobile ions, but there are no charges available for conduction.

Figure 3–7(b) shows the two charge depletion regions of opposite sign on the two sides of the junction. Since the numerical values of the charge on the two types of ion are the same, for overall charge neutrality

$$N_d X_2 = N_a X_1 \tag{3–4}$$

The charge depletion region in each material is therefore inversely proportional to the donor or acceptor density in the material. The *effective width* of the junction (that is, the region over which the bulk properties of the *p*- or *n*-type material no longer apply) is $X_1 + X_2$. This quantity is of considerable importance in junction transistors as will be shown later. Also of interest in

both *p-n* junctions and transistors is the capacitance across the junction because of the two space charge regions (see Sec. 3–4).

There is an important relationship, known as *Poisson's equation*, which may be used to find the way in which the potential varies with distance in a

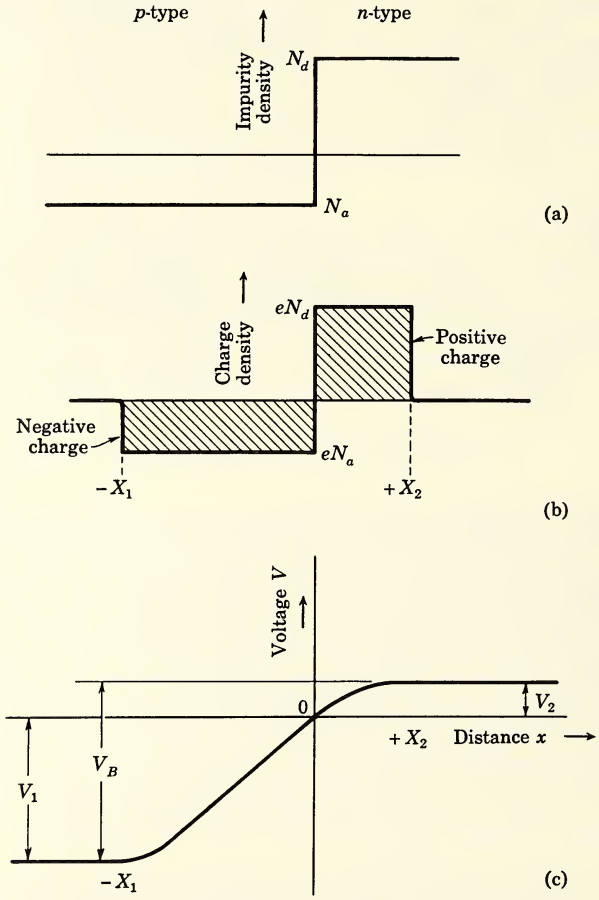


FIG. 3–7. (a) Impurity density. (b) Charge density. (c) Potential across a *p–n* junction.

region containing charge. In three dimensions, and written in nonvectorial form, Poisson's equation is

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = \frac{-\rho}{\varepsilon}$$

where ρ is the volume charge density in the material of permittivity ε.

In the more simple one-dimensional case that we are using in this analysis, Poisson's equation reduces to

$$\frac{d^2V}{dx^2} = \frac{-\rho}{\varepsilon} \tag{3-5}$$

In the space charge region of the p-type material in the junction shown in Fig. 3–7 and assuming that the junction has unit area,

$$\rho = -eN_a \tag{3-6}$$

Thus,

$$\frac{d^2V}{dx^2} = \frac{-\rho}{\varepsilon} = \frac{eN_a}{\varepsilon} \tag{3-7}$$

Integrating,

$$\frac{dV}{dx} = \frac{eN_a}{\varepsilon} x + C \tag{3-8}$$

and

$$V = \frac{eN_a x^2}{2\varepsilon} + Cx + D \tag{3-9}$$

Equation (3–9) applies only in the space charge region of the p-type material. To evaluate the constants, C and D, we will investigate the boundary conditions at the two ends of the region. At the right-hand side of this region in Fig. 3–7, we can write $V = 0$ when $x = 0$, if all voltages are measured by reference to the potential at the metallurgical boundary of the two materials. In the bulk p-type semiconducting material, we know that the potential is uniform. Therefore, in Fig. 3–7, the space charge region can be said to end at a point $x = -X_1$ where $dV/dx = 0$. Inserting these two boundary conditions into Eqs. (3–9) and (3–8), respectively,

$$D = 0$$

from Eq. (3–9), and

$$C = \frac{eN_a}{\varepsilon} X_1$$

from Eq. (3–8). If $V = V_1$ when $x = -X_1$,

$$V_1 = \frac{-eN_a}{2\varepsilon} X_1^2 \tag{3-10}$$

If Poisson's equation for the n-type depletion region is considered,

$$\frac{d^2V}{dx^2} = eN_d$$

and by a similar argument it can be shown that

$$V_2 = \frac{eN_d}{2\varepsilon} X_2^2 \tag{3-11}$$

where $V = V_2$ at $x = X_2$ (the limit of the n-type depletion region in Fig. 3–7).

The total potential across the junction when no external voltage is supplied is the contact potential, V_B.

$$V_B = |V_1| + |V_2| = \frac{e}{2\varepsilon}(N_a X_1^2 + N_d X_2^2) \tag{3–12}$$

From charge neutrality,

$$N_a X_1 = N_d X_2 \tag{3–4}$$

Eliminating X_2 from Eqs. (3–12) and (3–4),

$$V_B = \frac{eN_a X_1^2}{2\varepsilon}\left(1 + \frac{N_a}{N_d}\right)$$

or

$$X_1 = \left[\frac{2\varepsilon V_B}{eN_a\left(1 + \frac{N_a}{N_d}\right)}\right]^{1/2} \tag{3–13}$$

Alternatively, if X_1 is eliminated,

$$X_2 = \left[\frac{2\varepsilon V_B}{eN_d\left(1 + \frac{N_d}{N_a}\right)}\right]^{1/2} \tag{3–14}$$

Equations (3–13) and (3–14) show that the depletion widths can be calculated from a knowledge of the contact potential and the donor and acceptor densities in the *n*- and *p*-type semiconducting regions.

It is possible to calculate the contact potential if N_d and N_a are given for the two regions. For the junction of Fig. 3–7, we can write n_n for the equilibrium density of electrons in the *n*-type bulk material (away from the space charge region), and n_p for the equilibrium density of electrons in the *p*-type bulk material (away from the space charge region). Then $n_n = N_d$ where N_d refers to the *n*-type material, and $n_p = n_i^2/N_a$ where N_a refers to the *p*-type material, and n_p is obtained from the law of mass action [Eq. (2–34)]. In Fig. 3–8, we have repeated the energy band diagram of the *p*-*n* junction. The Fermi levels on the two sides of the junction are coincident, and the diagram also indicates the relative positions of the Fermi levels of intrinsic material with reference to the energy gap. If E_p and E_n are the energy levels of the bottom of the conduction band on the *p*-type and *n*-type side respectively, we can write

$$E_p - E_n = eV_B \tag{3–15}$$

In addition, from Eqs. (2–23) and (2–25),

$$n_n = N_d = n_i \exp\left(\frac{\Delta E}{kT}\right)$$

and

$$n_p = \frac{n_i^2}{N_a} = n_i \exp\left(\frac{-\Delta E'}{kT}\right)$$

where ΔE and $\Delta E'$ are as defined in Secs. 2–4 and 2–5 and as shown in Fig. 3–8. From the diagram,

$$\Delta E + \Delta E' = eV_B \tag{3-16}$$

Thus,

$$\frac{n_n}{n_p} = \exp\left(\frac{\Delta E + \Delta E'}{kT}\right) = \exp\left(\frac{eV_B}{kT}\right) \tag{3-17}$$

or

$$V_B = \frac{kT}{e}\log_e\left(\frac{n_n}{n_p}\right) = \frac{kT}{e}\log_e\left(\frac{N_d N_a}{n_i^2}\right) \tag{3-18}$$

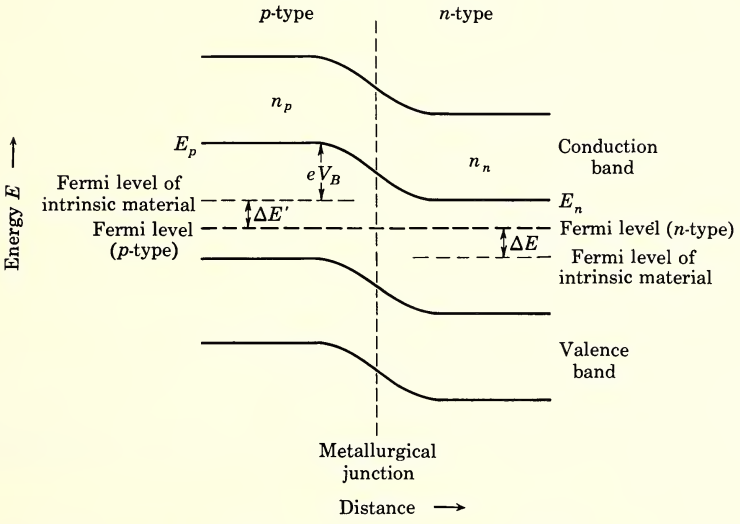


FIG. 3–8. Energy band diagram for the p–n junction.

Example:

A p-n junction is formed from germanium of resistivity 1 ohm cm (p side) and 0.1 ohm cm (n side). When operating at room temperature (300° K), what is the voltage across the junction (contact potential), and what is the total width of the depletion region? [*Note*: Many of the quantities required in this calculation are commonly given in c.g.s. units whereas most electrical calculations are made using practical (M.K.S.) units. Familiarity must be obtained in converting from one set of units to the other.]

Since the resistivity of both sides is considerably less than the resistivity of intrinsic germanium, $N_d \gg n_i$ (n-type) and $N_a \gg n_i$ (p-type). For the n-type material where $\sigma_n = 10$ mho cm^{-1}, $\sigma_n = N_d e\mu_n$ and the electron mobility, μ_n, from Fig. 1–21 is 3000 cm^2 volt^{-1} sec^{-1}. Hence,

$$N_d = \frac{\sigma_n}{e\mu_n} = \frac{10}{(1.6 \times 10^{-19})3000} \approx 2.1 \times 10^{16}\ \text{cm}^{-3}$$

$$= 2.1 \times 10^{22}\ \text{m}^{-3}$$

For the p-type material, the hole mobility, μ_p, from Fig. 1–22 is 1650 cm^2 $volt^{-1}$ sec^{-1} when $\sigma_p = 1$ mho cm^{-1}, and

$$N_a = \frac{\sigma_p}{e\mu_p} = \frac{1}{(1.6 \times 10^{-19})1650} = 3.8 \times 10^{15}\ cm^{-3}$$
$$= 3.8 \times 10^{21}\ m^{-3}$$

At a temperature of 300° K,

$$\frac{kT}{e} = \frac{(1.38 \times 10^{-23})300}{1.6 \times 10^{-19}} = 25.8\ mv$$

Hence,

$$V_B = \frac{kT}{e} \log_e \left(\frac{N_d N_a}{n_i^2}\right)$$
$$= 25.8 \log_e \left[\frac{(2.1 \times 10^{16})3.8 \times 10^{15}}{6.25 \times 10^{26}}\right]$$
$$= 303\ mv$$

From Eq. (3–13),

$$X_1 = \left[\frac{2\varepsilon V_B}{eN_a\left(1 + \frac{N_a}{N_d}\right)}\right]^{1/2}$$

Working in M.K.S. units and noting that the relative permittivity of germanium is 16,

$$X_1 = \left[\frac{2(16)(8.85 \times 10^{-12})0.303}{1.6 \times 10^{-19}(3.8 \times 10^{21})(1 + 0.183)}\right]^{1/2} = 3.45 \times 10^{-7}\ m$$

Similarly, from Eq. (3–14),

$$X_2 = \left[\frac{2(16)(8.85 \times 10^{-12})0.303}{1.6 \times 10^{-19}(2.1 \times 10^{22})(1 + 5.53)}\right]^{1/2} = 6.25 \times 10^{-8}\ m$$

$$\text{Total depletion width} = |X_1| + |X_2| = 4.08 \times 10^{-7}\ m$$

Because of the contact potential across a p-n junction, the energy of an electron at the bottom of the conduction band in the p-type material is greater than that of a corresponding electron at the bottom of the conduction band in the n-type material by an amount eV_B (Fig. 3–8). The presence of the potential across the p-n junction, therefore, can be thought of as a *potential barrier* existing between the two sides. The voltage, V_B, is in such a direction as to present a potential barrier to electrons which try to move from the n to the p side, whereas, it will assist electrons moving from the p to the n side. For an open-circuited junction, the net flow of current must be zero, and this condition is used in the analysis of the junction in Sec. 3–5.

3–4 Semiconductor Junction with Applied Voltage: Junction Capacitance

When a voltage is applied to a *p-n* semiconductor junction, it is found that the voltage-current characteristic is nonlinear. If the *p*-type material is made positive with respect to the *n*-type, the junction is said to be *forward biased*

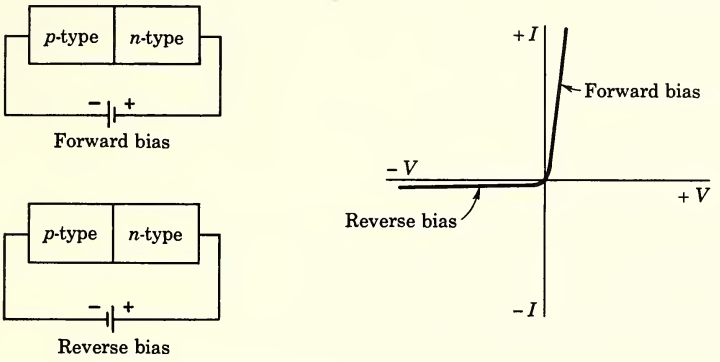


FIG. 3–9. The *p–n* junction and its characteristics under forward and reverse bias.

since this is the direction in which the resistance of the device is low and the current flows more readily. When the *n*-type is made positive with respect to the *p*-type, the junction is *reverse biased*, and its resistance is high (Fig. 3–9).

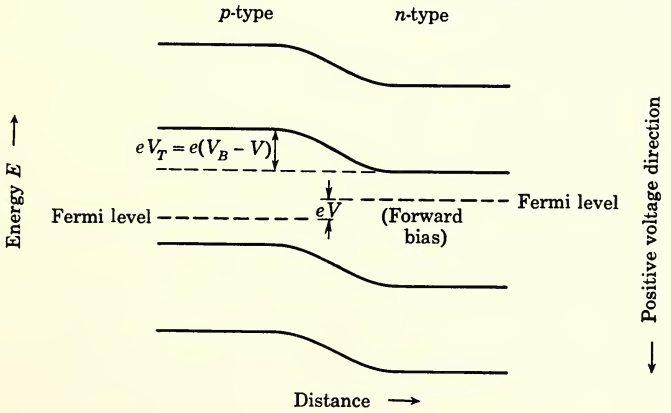


FIG. 3–10. A *p–n* junction with forward bias.

This rectifying property of the *p-n* junction will later be explained in terms of the previously developed theory of semiconductors.

When only a small current is flowing in a typical *p-n* junction, the voltage drop across the *p*- or *n*-type bulk material is usually small compared with the

voltage drop across the junction. In the analysis that follows, therefore, the voltage drop in the semiconducting material itself will be neglected, and all the applied voltage will be assumed to be dropped across the junction.

The presence of an applied voltage across the junction adds to, or subtracts from, the barrier voltage V_B. Consider the *p-n* junction with forward bias as shown in Fig. 3–10. When an electron passes across the junction from the *n*- to the *p*-type material, it is moving in the same direction as the force due to the applied electric field. Thus the potential energy in the *p*-type material is less than it would have been in an unbiased junction by an amount V electron volts, and in Fig. 3–10, the conduction and valence bands have been lowered by this value. (Compare Fig. 3–10 with the energy bands for the unbiased junction given in Fig. 3–8.) The energy band structure of the

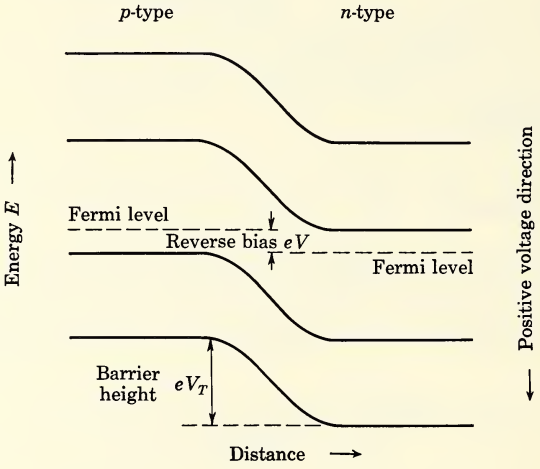


FIG. 3–11. A *p–n* junction with reverse bias.

p-type material is unchanged, however, and so the Fermi level, too, is lowered V electron volts with respect to its counterpart in the *n*-type material.

Since the unbiased barrier voltage (the contact potential V_B) is reduced by a forward bias V, the barrier voltage in the general case is given by V_T where

$$V_T = V_B - V \tag{3–19}$$

Figure 3–11 illustrates the case of the *p-n* junction with reverse bias. The barrier height is now increased over the unbiased value.

Referring back to the analysis of the preceding section, Eqs. (3–4) to (3–14) still apply if the total barrier voltage V_T is substituted in place of the unbiased value V_B. For example, Eq. (3–13) can be rewritten as

$$X_1 = \left[\frac{2\varepsilon V_T}{eN_a\left(1 + \dfrac{N_a}{N_d}\right)}\right]^{1/2} \tag{3–20}$$

where X_1 is now the depletion width in the p-type material with an applied voltage V. Similarly,

$$X_2 = \left[\frac{2\varepsilon V_T}{eN_d\left(1 + \frac{N_d}{N_a}\right)}\right]^{1/2} \tag{3–21}$$

where X_2 is the depletion width in the n-type material with applied voltage V.

Each space charge region of the junction contains a charge which is numerically equal to Q, where

$$|Q| = eN_aX_1 = eN_dX_2 = \left[\frac{2e\varepsilon N_aN_dV_T}{(N_a + N_d)}\right]^{1/2}$$

from Eqs. (3–20) and (3–21) and from Fig. 3–7(b). Equation (3–19) shows that, when a forward voltage V is applied, V_T decreases, and Q is diminished. From Fig. 3–7(b), the n-type region contains positive charge, and the p-type region contains negative charge. These charges are both of the opposite sign to the charges on an equivalent capacitor when a forward bias is applied. Thus the junction charge, Q, should be written with a negative sign to give

$$Q = -\left[\frac{2e\varepsilon N_aN_dV_T}{(N_a + N_d)}\right]^{1/2} \tag{3–22}$$

The variation in junction charge, Q, as the applied voltage, V, is altered constitutes an *incremental junction capacitance*, C, which is given by $C = dQ/dV$. From Eq. (3–19),

$$dV_T = -dV$$

and so, by differentiation of Eq. (3–22),

$$C = \frac{dQ}{dV} = \left[\frac{e\varepsilon N_aN_d}{2(N_a + N_d)}\right]^{1/2} V^{-1/2} \tag{3–23}$$

By an algebraic manipulation of Eqs. (3–20), (3–21), and (3–23), we can write,

$$C = \frac{\varepsilon}{|X_1| + |X_2|} \tag{3–24}$$

The last two equations show that across an abrupt semiconductor junction, there is a *nonlinear incremental capacitance* of value equal to the quotient of the permittivity of the medium and the total depletion width. For the junction used as an example in Sec. 3–3,

$$|X_1| + |X_2| = 4.08 \times 10^{-7}\ \text{m}$$

and, taking the relative permittivity of germanium as 16, the capacitance per unit area is

$$C = \frac{\varepsilon}{|X_1| + |X_2|} = \frac{16(8.85 \times 10^{-12})}{4.08 \times 10^{-7}} = 346\ \mu\text{f m}^{-2}$$

For a junction of area 10^{-6} m^2, $C = 346$ pf. This is the capacitance across the junction when the applied voltage is small compared with V_B. In practice, the capacitance is more frequently of concern when the *p-n* junction is reverse biased, since then its reactance may be smaller than the resistance of the junction in the reverse direction. If the applied voltage is -10 volts,

$$V_T = V_B - V = 0.303 + 10 = 10.3 \text{ volts}$$

The capacitance of the junction then becomes 59 pf. Note from Eqs. (3–23) and (3–24) that if one of the regions is of high conductivity, it has little effect on the junction capacitance which is then determined by the conductivity, i.e., the doping, of the other region.

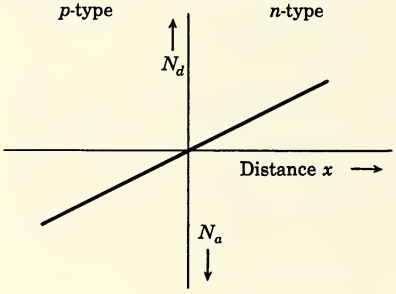


FIG. 3–12. A linearly graded junction where $N_d - N_a = ax$ and a is a constant.

For a linearly graded junction where the donor and acceptor densities are as given in Fig. 3–12, the incremental junction capacitance is proportional to $V_T^{-1/3}$. The relationship between capacitance and voltage can be derived in a similar way to the analysis for the abrupt junction (see problem 3–4). The nonlinear capacitance versus voltage characteristic is useful in some types of electronic circuitry.

3–5 Flow of Current across a *p-n* Junction: The Rectifier Equation

In this section, we shall quantitatively discuss the flow of current across a *p-n* junction and derive the form of the rectifier equation. This will give a physical picture of the semiconductor rectifier before the more exact analysis is given in Chap. 4.

Consider Fig. 3–13 which shows the energy bands of a *p-n* junction. In the diagram, a forward bias has been applied, and so the Fermi level on the *n*-type side, E_{Fn}, is higher than that on the *p*-type side by the value eV. There are four possibilities for current flow across the junction, and these have been shown in the figure. In the valence band, holes can flow to the right or to the left, and the two current densities have been separately identified as J_1^+ and J_2^+ respectively. In the conduction band, electrons can flow to the right or to the left. A flow of electrons to the left produces a conventional current flow to the right since electrons carry a negative charge. For this reason, the arrows in the conduction band on Fig. 3–13 indicate the direction of flow of the electrons, and a minus sign has been placed before the current densities. Thus current density J_1^- is flowing to the right, and current density J_2^- is flowing to the left. This means that the two current densities J_1^-

and J_1^+ add together in the right-hand direction and current densities J_2^- and J_2^+ add together in the left-hand direction. The net current density flowing toward the right is

$$J = (J_1^- + J_1^+) - (J_2^- + J_2^+) \tag{3–25}$$

The variation of all four current densities with applied voltage, V, will now be discussed.

In the conduction band, current density J_2^- is caused by a flow of electrons from the p-type material down the potential hill into the n-type. In a typical junction, this is a small current of a few microamps since its supply of charge

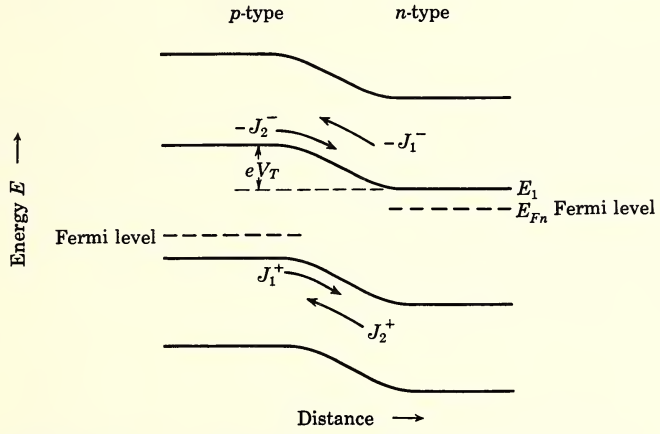


FIG. 3–13. A p–n junction showing electron and hole current densities. (Note that the arrows show the direction of flow of both electrons and holes. For J_1^- and J_2^- the flow of conventional current is in the opposite direction.)

is the conduction band of the p-type material where electrons are scarce. The mechanism of the transfer of charge from the bulk material to the depletion region is mainly that of diffusion which will be discussed in Chap. 4. Here, we note only that the passage of electrons into the depletion region by diffusion is nearly independent of the voltage across the junction since the two regions are substantially independent of one another. Once an electron is at the edge of the potential barrier (coordinate $-X_1$ on Fig. 3–7), it will be swept across the junction by the action of the electric field. Changing the height of the barrier will not alter J_2^- provided the barrier voltage remains in the sense shown in Fig. 3–13. In other words, J_2^- will be constant provided the electrons flow down a potential hill. This condition is true for all values of reverse bias, and also when the forward bias is less than V_B. J_2^- is often known as a *saturation current density* for this reason.

J_1^- is the other current density in the conduction band, and here electrons flow from the n-type to the p-type semiconductor. The conduction band of

the n-type material provides a copious source of electrons for this current. Provided the current is not excessive, the only restriction on the flow of the current is that of the presence of the barrier potential V_T. The probability of occupation of the states in the conduction band of the n-type material decreases exponentially with increasing energy in the band according to $\exp[(E_{Fn} - E)/kT]$. For a potential barrier of height V_T above the bottom of the conduction band of the n-type material (as shown in Fig. 3–13), the number of electrons which are an available source of current is proportional to $\exp(-eV_T/kT)$. Thus the flow of electrons from right to left across the barrier causes a current

$$J_1^- = C_1 \exp\left(\frac{-eV_T}{kT}\right) \tag{3–26}$$

where C_1 is a constant to be determined.

A similar argument may be used to find the two positive hole current densities in the valence band. J_2^+ is a saturation current density which does not depend on the height of the potential barrier. Since holes carry a positive sign, holes flowing upwards in Fig. 3–13 are effectively flowing down the potential barrier. Current density J_1^+, which is flowing in the opposite direction, is moving up the potential barrier and is given by

$$J_1^+ = C_2 \exp\left(\frac{-eV_T}{kT}\right) \tag{3–27}$$

where C_2 is another constant.

There is no net current flow across the junction when the applied voltage, V, is zero. From Eq. (3–19), $V_T = V_B$, and

$$J_1^- - J_2^- = 0 \tag{3–28}$$

and

$$J_1^+ - J_2^+ = 0 \tag{3–29}$$

because current flow in one band is independent of the flow in the other band and charge cannot be accumulated in either band. Substituting for J_1 from Eq. (3–26) in Eq. (3–28) and using $V_T = V_B$

$$C_1 \exp\left(\frac{-eV_B}{kT}\right) - J_2^- = 0$$

or

$$C_1 = J_2^- \exp\left(\frac{eV_B}{kT}\right) \tag{3–30}$$

Similarly,

$$C_2 = J_2^+ \exp\left(\frac{eV_B}{kT}\right) \tag{3–31}$$

Equations (3–26) and (3–30) relate J_1^- and J_2^- in terms of V_T and V_B. Similarly, Eqs. (3–27) and (3–31) give the relationship between J_1^+ and J_2^+. For the

general case of $V \neq 0$, we may substitute for J_1^- and J_1^+ in Eq. (3–25), to give

$$\begin{aligned} J &= (J_1^- + J_1^+) - (J_2^- + J_2^+) \\ &= J_2^- \exp\left[\frac{e(V_B - V_T)}{kT}\right] + J_2^+ \exp\left[\frac{e(V_B - V_T)}{kT}\right] - (J_2^- + J_2^+) \\ &= (J_2^- + J_2^+)\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \\ &= J_2\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \end{aligned} \tag{3–32}$$

where $J_2 = J_2^- + J_2^+$ (the sum of the saturation current densities), and $V = V_B - V_T$ from Eq. (3–19).

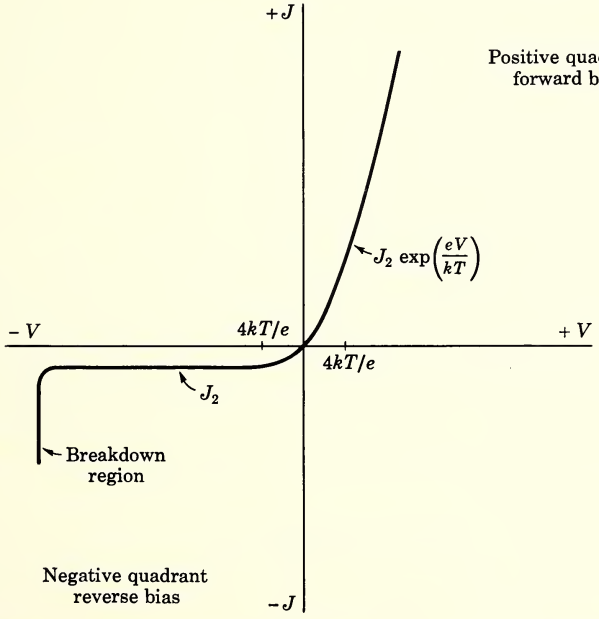


FIG. 3–14. The voltage-current characteristic of a p–n junction rectifier.

Equation (3–32) is known as the rectifier equation and is plotted in Fig. 3–14. In the positive quadrant, when a forward bias is applied of magnitude $eV > 4kT$, $\exp(eV/kT) \gg 1$, and $J = J_2 \exp(eV/kT)$. In the negative quadrant, when a reverse bias of magnitude $-eV > 4kT$ is applied, $\exp(eV/kT) \ll 1$, and $J = J_2$. At room temperature ($T = 300°$ K), $4kT = 103$ millielectron volts.

When the junction is biased in the forward direction, the potential barrier, V_T, is lowered, and the currents J_1^- and J_1^+ are increased. For the reverse-biased junction, the potential barrier is increased, and the saturation current

densities J_2^- and J_2^+ dominate the expression. The value of the saturation current is derived in Chap. 4.

As the forward current increases, the characteristic in Fig. 3–14 becomes more linear. This is because the ohmic resistance of the semiconductor (which has so far been neglected) causes a significant voltage drop to occur in the body of the semiconductor.

3–6 The Breakdown Region of a *p-n* Junction

As the voltage is increased in the negative direction in Fig. 3–14, we come to a part of the curve labelled "breakdown region." Here, the current increases very rapidly. This is the effect of *avalanche breakdown*. An electron acquires sufficient energy in its passage in the electric field that it can excite an electron from the valence band to the conduction band, and so add an electron-hole pair to the existing charge carriers. The new charge carriers now take place in a secondary collision process of the same nature, and an avalanche of charges is obtained. This mechanism is very similar to the Townsend avalanche which is of great importance in gas discharge phenomena.

The name "Zener breakdown" has been widely used to describe the breakdown region of the *p-n* rectifier characteristic and is still in use even though the avalanche mechanism is usually operative. Zener breakdown occurs when the electric field is so high that electrons are pulled out of their covalent bonds in a semiconductor. For most *p-n* junctions, avalanche breakdown will occur first. The avalanche breakdown voltage is very stable and forms a useful reference voltage in electronic circuitry. Semiconductor manufacturers produce special devices for this purpose known as "Zener" (a misnomer), voltage regulator, or avalanche diodes.

BIBLIOGRAPHY

Greiner, R. A., *Semiconductor Devices and Applications*, New York: McGraw-Hill Book Company, Inc., 1961

Middlebrook, R. D., *An Introduction to Junction Transistor Theory*, New York: John Wiley & Sons, Inc., 1957

Shockley, William, *Electrons and Holes in Semiconductors*, New York: D. Van Nostrand Company, Inc., 1950

Smith, R. A., *Semiconductors*, London: Cambridge University Press, 1959

Spangenburg, Karl R., *Fundamentals of Electron Devices*, New York: McGraw-Hill Book Company, Inc., 1957

Sproull, Robert L., *Modern Physics*, New York: John Wiley & Sons, Inc., 1956

PROBLEMS

3–1 Explain the terms (a) valence, (b) work function, (c) contact potential with reference to metals or metal junctions.

3–2 An abrupt *p-n* junction is made of silicon where the resistivities of the two sides are 2 ohm cm (*p*-side) and 1 ohm cm (*n*-side). Using the information given in Secs. 1–18 and 3–3 compute the contact potential and the total width of the depletion regions for zero applied voltage.

ans. 0·30 volts, 3.1×10^{-7} m

3–3 Using the example in Sec. 3–3, find the positions of the Fermi levels on the two sides of the junction at 25° K. Using Eq. (2–35) and assuming that the exponential term dominates the expression, show that the variation in Fermi level with temperature for the *n*-type material may be calculated. From this argument, show that the contact potential decreases to zero as the temperature is raised.

3–4 Show that the incremental capacitance of a linearly graded junction is given by

$$C = \left(\frac{ea\varepsilon^2}{12V_T}\right)^{1/3}$$

where $N_d - N_a = ax$ and a is a constant (see Fig. 3–12).

3–5 What is the incremental junction capacitance of a germanium *p-n* junction of area 10^{-3} cm^2 when a reverse voltage of 1 volt is applied? The conductivity of the *n*-type side is 1 mho cm^{-1} and that of the *p*-type side is 2 mho cm^{-1}.

3–6 What is meant by a space charge region? Discuss the variation in the space charge region of an abrupt *p-n* junction which occurs when the applied voltage is changed. Show that the junction has an incremental capacitance and suggest a use for this nonlinear capacitance.

3–7 Show that for the condition where the acceptor density in the *p*-type region is greater than the donor density in the *n*-type region, the total depletion width of the *p-n* junction when there is no applied voltage is

$$\left(\frac{2\varepsilon\mu_n V_B}{\sigma_n}\right)^{1/2}$$

3–8 Find the depletion widths and the junction capacitance for the *p-n* junction used as an example in Secs. 3–3 and 3–4 when (a) a reverse bias of 4 volts is applied (b) when a forward bias of 0.04 volts is applied.

3–9 Describe the four currents which may be considered to flow across the *p-n* junction. What is meant by the saturation current density? Show that for very small values of applied voltage (whether forward or

reverse) the *p-n* junction behaves as a small linear resistance of value $0.026/I_2$ ohms at room temperature, where I_2 is the total saturation current across the junction.

3–10 A *p-n* junction is constructed with conductivities of 200 mhos m^{-1} (*n*-side) and 500 mhos m^{-1} (*p*-side). The total length of the device is 2 mm and the cross-sectional area is 1 sq mm. The saturation current in the reverse direction is 2 microamps at room temperature. Compute the voltage across the whole device when the current is 10 ma assuming the junction is at the center of the device and that no temperature rise takes place.

4

The Continuity Equation

It has been shown that the number of free electrons and positive holes available for the passage of current in a semiconductor is considerably less than the number of free electrons in a metal. This fact is of importance in analyzing the flow of current in semiconducting materials since the equilibrium charge density is disturbed by the passage of a comparatively small current density. In this chapter, a fundamental equation is developed which governs the behavior of charge carriers in the material. This is known as the *continuity equation.*

4–1 Derivation of the Continuity Equation

The analysis of the charge distributions and current flow in a piece of semiconducting material will be carried out by taking into account the effects of three phenomena. These are:

1. the generation and recombination of electrons and positive holes,
2. the drift of charge in an electric field,
3. the diffusion of charge as a result of a charge concentration gradient.

Among the other effects which may have to be considered in particular cases are temperature gradients and applied magnetic fields. However, in most of the problems which will be encountered in practice it is necessary only to consider the three most important phenomena listed above.

The continuity equation is formed by adding the three rates of delivery or removal of charge in a given volume of the semiconductor and equating this

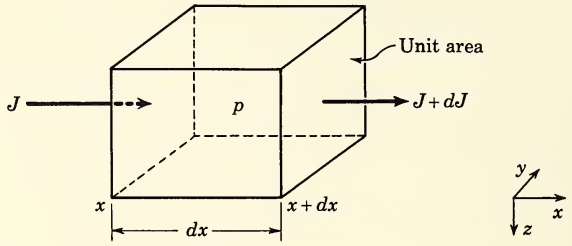


FIG. 4–1. Volume element used in the derivation of the continuity equation.

total to the rate of change of charge density in the material. This computation is a more sophisticated version of the calculation of the change in height of liquid in a tub when water is entering through an inlet and leaving through an outlet. The semiconductor problem is complicated by having two charge carriers of opposite sign (negative electrons and positive holes) and by the possibility of the presence of a nonuniform charge density through the material. In practice, two continuity equations may be established: one will be for electrons, and one will be for positive holes.

The continuity equation is extremely important in the understanding of the properties of semiconductor diodes and transistors, and it will be used extensively in this book. For simplicity, the one-dimensional form of the equation will be derived here. Although more complete, the three-dimensional form is more difficult to solve and gives little additional information for our purpose.

The volume element in the semiconductor which is to be used in the derivation of the continuity equation is shown in Fig. 4–1. It is a three-dimensional element of unit area in the yz plane and of length dx in the x

direction. The total volume of the element is dx. The continuity equation will first be derived for positive holes only.

Let the average density of positive holes in the volume element at any time be p per unit volume. Restricting the argument to the one-dimensional case, assume that the current density entering the element by the left-hand face is J, and the current density leaving the right-hand face is $J + dJ$. The difference between these two current densities shows that dJ/e positive holes are being removed from the volume element every second by current flow. Note that this lack of continuity in current indicates that other mechanisms are operating in the delivery, removal, or storage of charge in the element.

The three phenomena which affect the charge density in the semiconductor will be discussed in the order in which they were listed.

1. *Generation and recombination of electrons and positive holes.* It was shown in Sec. 1–20 that the rate of change of electron density with time could be represented by

$$\frac{dn}{dt} = \frac{n_0 - n}{\tau_n}$$

where the generation rate, n_0/τ_n, and the recombination rate, n/τ_n, were the only factors affecting the electron density.

By a similar argument,

$$\frac{dp}{dt} = \frac{p_0 - p}{\tau_p} \tag{4–1}$$

where p_0 is the equilibrium density of positive holes in the material, p is the positive hole density at any time, and τ_p is the hole lifetime. This equation determines the change in hole density because of generation and recombination.

2. *Drift of charge in an electric field.* Suppose an electric field, $\mathscr{E}$, is applied along the x axis of the volume element. Positive holes will flow in the x direction with an average velocity v where

$$v = \mu_p \mathscr{E} \tag{4–2}$$

and μ_p is the mobility of positive holes in the material. This flow of charge constitutes a current density in the x direction of magnitude

$$J_{\mathrm{Dr}}^{+} = pev = pe\mu_p \mathscr{E} \tag{4–3}$$

where the superscript + indicates that this is a current of positive holes, and the subscript Dr stands for drift to distinguish it from the case of diffusion. Equation (4–3) will be used later when the terms in the continuity equation are assembled.

3. *Diffusion of charge.* The phenomenon of *diffusion* plays an important role in transistor operation. Diffusion is a process whereby charge carriers (holes

or electrons) can move in a material *independently of the electric field.* Diffusion is the flow of carriers from regions of high carrier density to regions of low density and, therefore, occurs only where there is nonuniform carrier density in the material. As an illustration, assume that the hole density at one point in a semiconductor specimen is increased by some means. There will then be a flow of holes from the high to the low density region so that the hole density becomes more uniform in the material. If the mechanism which produced the excess holes stops operating, the effect of diffusion is to bring the hole density back to uniformity in the specimen. At all times, diffusion is a flow of carriers in the direction of decreasing carrier concentration regardless of the sign of the charge on the carrier.

We can represent the diffusion of holes in a semiconductor by the equation,

$$N = -D_p \frac{dp}{dx} \tag{4–4}$$

In this equation we are postulating a change in hole density in the x direction along the volume element and stating that N is the number of holes diffusing per unit area in the x direction per second. The minus sign is included in the equation because the diffusion of holes is in the direction of decreasing concentration, that is, N will be positive when dp/dx is negative. The diffusion constant D_p is a constant for positive holes in a particular material. A typical value for D_p is 49 cm^2 sec^{-1} for intrinsic germanium (see Table 1–2).

The current density carried by the diffusing holes is

$$J_{\mathrm{Di}}^{+} = eN = -eD_p \frac{dp}{dx} \tag{4–5}$$

The subscript Di indicates diffusion and the superscript + shows that the current is carried by positive holes. Here the negative sign shows that the current is diffusing away from the region of maximum concentration density.

The sum of the drift and diffusion currents must be taken into account in determining the rate of change of charge density with time. The total current density of positive holes is

$$J^{+} = J_{\mathrm{Dr}}^{+} + J_{\mathrm{Di}}^{+} \tag{4–6}$$

and the net flow of current from the element is dJ (see Fig. 4–1).

Considering only the phenomena of drift and diffusion, the charge density may be postulated to change at the rate of $\partial p/\partial t$. Thus the rate of increase of the total number of positive holes in the element is

$$\frac{\partial p}{\partial t} dx$$

since dx is the volume of the element. If the net current flow is responsible for

this change in holes contained in the element,

$$\frac{\partial p}{\partial t}\,dx = -\frac{dJ}{e}$$

or

$$\frac{\partial p}{\partial t} = -\frac{1}{e}\frac{\partial J}{\partial x} \tag{4–7}$$

where the negative sign appears because the net flow of current dJ was defined as being out of the volume element.

Equation (4–7) takes into account only drift and diffusion. When the generation and recombination processes are also included,

$$\frac{\partial p}{\partial t} = \frac{p_0 - p}{\tau_p} - \frac{1}{e}\frac{\partial J}{\partial x} \tag{4–8}$$

Substituting the values of the drift and diffusion currents from Eqs. (4–3), and (4–5) into Eq. (4–6), and differentiating with respect to x, Eq. (4–8) can be rewritten as

$$\frac{\partial p}{\partial t} = \frac{p_0 - p}{\tau_p} - \mu_p \frac{\partial}{\partial x}(p\mathscr{E}) + D_p \frac{\partial^2 p}{\partial x^2} \tag{4–9}$$

This is the *continuity equation for positive holes.*

The continuity equation for electrons may be derived by reconsidering the three terms separately. From Eq. (1–32), the rate of change of electron charge density is related to the generation and recombination rate by

$$\frac{dn}{dt} = \frac{n_0 - n}{\tau_n} \tag{4–10}$$

The drift of charge in an electric field $\mathscr{E}$ is given by

$$J_{\mathrm{Dr}}^{-} = ne\mu_n\mathscr{E} \tag{4–11}$$

where the superscript $-$ indicates electron flow. The diffusion of electrons in the x direction can be represented by a similar equation to Eq. (4–4), i.e.,

$$N = -D_n \frac{dn}{dx}$$

where N is now the number of electrons diffusing across unit area per second, and D_n is the diffusion constant for electrons. Since electrons have a negative charge, the current density carried by this flow of electrons per second per unit area is

$$J_{\mathrm{Di}}^{-} = -eN = eD_n \frac{dn}{dx} \tag{4–12}$$

When the drift and diffusion current densities are considered, we see that a net conventional current flow to the right is caused by electron motion to the

left. Thus the corresponding equation to Eq. (4–7) is

$$\frac{\partial n}{\partial t} = \frac{1}{e}\frac{\partial J}{\partial x}$$

and the *continuity equation for electrons* becomes

$$\frac{\partial n}{\partial t} = \frac{n_0 - n}{\tau_n} + \mu_n \frac{\partial}{\partial x}(n\mathscr{E}) + D_n \frac{\partial^2 n}{\partial x^2} \tag{4–13}$$

4–2 Application of the Continuity Equations to the Abrupt *p-n* Junction Operating with a Constant Current

The continuity equations may be solved under various conditions of operation of the *p-n* junction. A simple, yet most important, case is that of a constant current flowing across a junction, since the solution of this problem leads to an understanding of the junction transistor. When the total junction current density is a constant, the hole and electron densities in the two regions are also constant with respect to time. Thus, wherever the continuity equations are applied, we can put

$$\frac{\partial p}{\partial t} = \frac{\partial n}{\partial t} = 0$$

The continuity equations for holes and electrons can now be written as

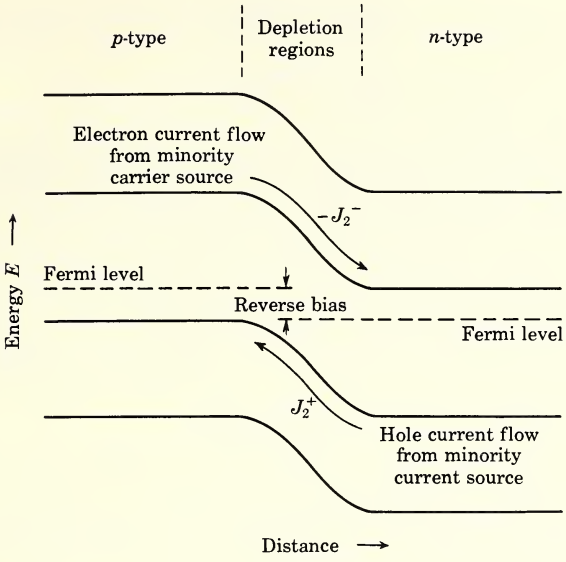
$$0 = \frac{p_0 - p}{\tau_p} - \mu_p \frac{d(p\mathscr{E})}{dx} + D_p \frac{d^2 p}{dx^2} \tag{4–14}$$

and

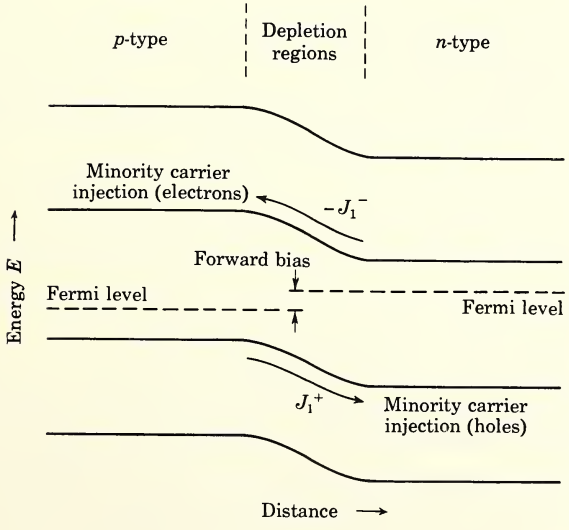
$$0 = \frac{n_0 - n}{\tau_n} + \mu_n \frac{d(n\mathscr{E})}{dx} + D_n \frac{d^2 n}{dx^2} \tag{4–15}$$

In Sec. 3–5 there was a discussion of the four current densities which can flow across the junction. When the junction is reverse biased so that $-eV > 4kT$, saturation currents J_2^- and J_2^+ are flowing as shown in Fig. 4–2(a). Both currents flow from regions where the respective charges are minority carriers and enter regions where they add to the majority carrier density. The total reverse saturation current is small so the majority carrier densities are hardly changed by the incoming charges, but the regions which supply the charge are greatly affected by the loss of their minority carriers. Close to the depletion regions on both sides of the junction, the minority carrier densities drop because the saturation currents are removing minority charges. Therefore, the continuity equations will be solved only for the minority charges close to the junction as these are the cases of maximum interest.

When the junction is forward biased with $eV > 4kT$, the two saturation currents can be neglected in comparison with the forward currents J_1^- and J_1^+. As shown in Fig. 4–2(b), these forward currents transfer charges from the majority carrier regions to the minority carrier regions. The injected



(a)



(b)

FIG. 4–2. Diagram showing the carriers flowing across a p–n junction (a) under reverse bias, (b) with forward bias. Arrows indicate carrier flow direction.

minority carrier densities will alter the thermal minority carrier equilibrium close to the junction, and so the solution of the continuity equations for the minority carrier regions will again be important.

In the depletion regions, there are very few mobile charge carriers present, and so little recombination takes place. We can therefore assume that hole and electron currents are, separately, continuous across the depletion region. This means that it is necessary to solve the continuity equations only for the minority carriers on the two sides of the junction to gain a knowledge of the total junction current. In general, we will solve the hole continuity equation to find the current carried by holes at the edge of the *n*-type region, and we will solve the electron continuity equation to find the current carried by electrons at the edge of the *p*-type region. Assuming no recombination in the depletion regions, the sum of these two currents is the constant total current throughout the material and across the junction.

Care must be taken before such restricted conditions are applied to a practical case. When very small currents are flowing, recombination may be a significant factor. This is particularly noticeable in silicon where the intrinsic carrier densities are much lower than in germanium at the same temperature. In addition, when the continuity equations are solved only for the minority carrier regions, it is on the assumption that the majority carrier levels remain at their thermal equilibrium values even though a current is flowing. We will distinguish among three conditions of operation for *p-n* junctions. *Very small current operation* refers to the case where the junction is reverse biased or operated in the forward direction with a very small current so that recombination in the depletion region must be taken into account. *Small-current operation* occurs when the current density is higher than the previous case but not so large that the majority carrier density is disturbed. This applies to a typical junction operated with a forward current of a few milliamps. Under *large-current operation* of a *p-n* junction, such a high current density is flowing that the majority carrier density is affected and both continuity equations must be solved in both regions to get an accurate analysis of carrier density distributions. In most of this book, small-current operation will be assumed. However, all three cases are considered in Sec. 9–7 where the variation of alpha with emitter current is discussed.

Close to a typical *p-n* junction, it is found that the diffusion current density is much larger than the drift current density under the condition of small-current operation. This is because the minority carrier levels on both sides of the junction are substantially modified by the flow of current whether the junction is forward or reverse biased as shown in Fig. 4–2. In the following analysis in this chapter, we shall neglect the drift terms in the continuity equations by comparison with the diffusion terms. (The justification for this approximation can be seen by inserting typical values for all the quantities involved in the continuity equations.) Equations (4–14) and (4–15) then become

$$\frac{p - p_0}{\tau_p} = D_p \frac{d^2p}{dx^2} \tag{4–16}$$

and

$$\frac{n - n_0}{\tau_n} = D_n \frac{d^2n}{dx^2} \tag{4–17}$$

The solution of these two equations for a p-n junction can be found using the same coordinate system for the two equations and putting $x = 0$ at the metallurgical junction. This is the most logical method but great simplification can be achieved by using two different coordinate systems as shown in Fig. 4–3. For the n-type material, x is measured in the positive direction to the right, and $x = 0$ coincides with the edge of the n-type depletion region. In the p-type material, the x coordinate is similarly measured from the edge of

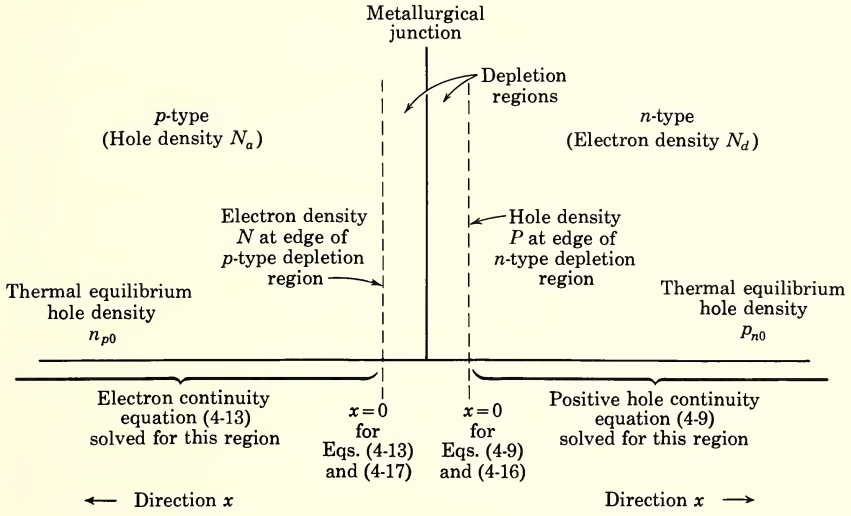


FIG. 4–3. Diagram showing the two coordinate systems used to simplify the solution of the two minority carrier continuity equations.

the p-type depletion region, but, here, the positive direction of x is to the left. The two equations are solved separately for the two regions and the situation will be clear if it is remembered that in each case, $x = 0$ signifies the edge of the depletion region and x is positive when moving away from the junction into the bulk material.

Consider first the solution of Eq. (4–16) in the n-type material. We can write Eq. (4–16) in the form,

$$\frac{d^2p_n}{dx^2} = \frac{p_n - p_{n0}}{D_p\tau_p} \tag{4–18}$$

where the additional subscript shows that the equation is restricted to the

n-type material. Since $D_p\tau_p$ can be considered constant, the solution is of the form

$$p_n - p_{n0} = K_1 \exp\left(\frac{-x}{\sqrt{D_p\tau_p}}\right) + K_2 \exp\left(\frac{x}{\sqrt{D_p\tau_p}}\right)$$

where K_1 and K_2 are constants of integration. This equation applies to the n-type material shown to the right-hand side of Fig. 4–3.

At a large distance from the junction, the positive hole density must approach its equilibrium value in the bulk material. Hence, as

$$x \to \infty, \qquad p_n \to p_{n0}$$

and so

$$0 = K_2 \exp\left(\frac{x}{\sqrt{D_p\tau_p}}\right)$$

giving

$$K_2 = 0$$

The solution of Eq. (4–18) is, therefore

$$p_n - p_{n0} = K_1 \exp\left(\frac{-x}{\sqrt{D_p\tau_p}}\right)$$

From Fig. 4–3, we can write $p_n = P$ at the edge of the depletion region where $x = 0$. Hence,

$$P - p_{n0} = K_1$$

and

$$p_n - p_{n0} = (P - p_{n0}) \exp\left(\frac{-x}{\sqrt{D_p\tau_p}}\right) \tag{4–19}$$

Equation (4–19) applies to both forward and reverse-biased junctions by inserting an appropriate value of P, as will be done later. For a change in x of value $\sqrt{D_p\tau_p}$, the charge density changes by a factor 1/exp because of recombination. The distance,

$$L_p = \sqrt{D_p\tau_p} \tag{4–20}$$

is known as the *diffusion length for positive holes* in this region, and it can be calculated from the values of the diffusion constant and the lifetime. Rewriting Eq. (4–19),

$$p_n - p_{n0} = (P - p_{n0}) \exp\left(\frac{-x}{L_p}\right) \tag{4–21}$$

in the n-type region. The application of this equation to junctions operating under reverse and forward bias is given in Secs. 4–4 and 4–5.

The continuity equation relating to electrons in the p-type material can be solved in a similar manner. From Eq. (4–17), we can write

$$\frac{d^2n_p}{dx^2} = \frac{n_p - n_{p0}}{L_n} \tag{4–22}$$

where the subscript p restricts the equation to the p-type material. The *diffusion length for electrons* in this region is L_n, where

$$L_n = \sqrt{D_n \tau_n}$$

and writing N for the electron density at the edge of the depletion region of the p-type material,

$$n_p - n_{p0} = (N - n_{p0}) \exp\left(\frac{-x}{L_n}\right) \tag{4–23}$$

by comparison with the argument in the previous paragraph. Equations (4–21) and (4–23) apply to reverse or forward-biased conditions of an abrupt p-n junction.

4–3 Calculation of the Charge Densities at the Edge of the Depletion Regions

We will have completed the solutions of the continuity equations when P and N are determined. Under conditions of small-current operation, the majority

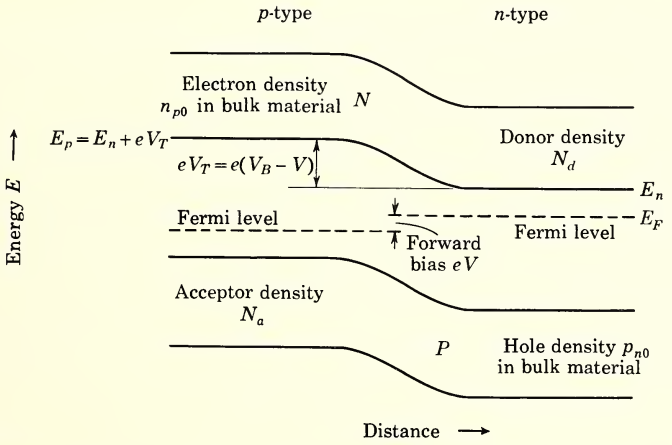


FIG. 4–4. Energy band diagram of a forward-biased p–n junction showing energy levels and carrier densities in the two regions.

carrier densities on both sides of the junctions are assumed to be undisturbed by the current flow. The minority current densities close to the depletion regions are determined by the donor and acceptor densities, and the Boltzmann factor as shown below.

The Boltzmann factor [Eq. (2–14)],

$$F(E) = \exp\left(\frac{E_F - E}{kT}\right) \tag{4–24}$$

is the probability, expressed in terms of the Fermi level, that an electron state at an energy E is occupied. Writing E_n as the lowest energy level in the conduction band of the n-type material and E_p as the lowest energy in the conduction band of the p-type material, and by reference to Fig. 4–4 and Eq. (3–19),

$$V_T = V_B - V \tag{4–25}$$

where V is the applied voltage across the junction, and

$$E_p = E_n + eV_T \tag{4–26}$$

If the applied voltage, V, were zero, the Fermi levels on the two sides of the junction would be in line. This means that the probabilities of occupation of states at a given energy on the two sides of the junction would be equal. In Fig. 4–4, however, a forward bias is shown. In this case, the Fermi level of the p-type side is eV below the Fermi level of the n-type side. For a given energy, E, in the conduction band, the ratio of the probability of occupancy in the n-type to the p-type is

$$\frac{\exp\left[(E_F - E)/kT\right]}{\exp\left[(E_F - eV - E)/kT\right]} = \exp\left(\frac{eV}{kT}\right) \tag{4–27}$$

where E_F is the Fermi level on the n-type side of the junction.

A current of electrons flows in the conduction band across the junction to equalize the probability of occupancy across the two sides. Since we have assumed small-current operation, this flow of charge does not disturb the majority carrier density in the n-type region but raises the minority carrier density in the p-type material by the factor $\exp(eV/kT)$. A small, continuous current is necessary to sustain this new minority carrier density on the left-hand side of the junction (p-type region) because diffusion and recombination are removing charge into the bulk material. Since the minority carrier density under thermal equilibrium in the bulk p-type material is n_{p0}, the increased electron density close to the depletion region is

$$N = n_{p0} \exp\left(\frac{eV}{kT}\right) \tag{4–28}$$

A similar argument can be applied to the valence band. The probability for the occurrence of holes has previously been given as

$$1 - F(E) \approx \exp\left(\frac{E - E_F}{kT}\right) \tag{4–29}$$

from Eq. (2–16), where E is now an energy lying within the limits of the valence band. The ratio of the probabilities of finding holes at an energy E in the valence band of the p-type and the n-type material is

$$\frac{\exp\left[(E - E_F)/kT\right]}{\exp\left[(E - E_F - eV)/kT\right]} = \exp\left(\frac{eV}{kT}\right) \tag{4–30}$$

as before. In this band, therefore, a current of positive holes moves from left to right to raise the minority carrier (hole) density at the edge of the n-type material by the factor $\exp(eV/kT)$. Since the hole density in the bulk n-type material is p_{n0},

$$P = p_{n0} \exp\left(\frac{eV}{kT}\right) \tag{4–31}$$

and the continuity equations can be written as

$$p_n - p_{n0} = p_{n0}\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \exp\left(\frac{-x}{L_p}\right) \tag{4–32}$$

$$n_p - n_{p0} = n_{p0}\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \exp\left(\frac{-x}{L_n}\right) \tag{4–33}$$

Care must be taken to apply these equations only to the regions where they were derived and only under small-current operation. Equations (4–32) and (4–33) apply when the junction is either reverse or forward biased since Eqs. (4–28) and (4–31) hold for negative or positive values of V, although we have described here the flow of current in the positive case only.

4–4 Solution of the Continuity Equations for a Reverse-biased Abrupt *p-n* Junction

The continuity equation solutions, Eqs. (4–32) and (4–33) apply directly to this case. Since the junction is reverse biased, V is negative, and $\exp(eV/kT) < 1$. Let us first examine the density of holes in the n-type region.

From Eq. (4–32), at the edge of the depletion region, $x = 0$, and so

$$p_n = P = p_{n0} \exp\left(\frac{eV}{kT}\right) < p_{n0} \tag{4–34}$$

Thus the minority charge density at the edge of the depletion region is less than the density in the bulk material. The physical reason for this is that the current J_2^+ is carrying holes across the junction into the p-type material, and these holes are being supplied from the bulk n-type material by diffusion. For electrons in the p-type material, at $x = 0$

$$n_p = N = n_{p0} \exp\left(\frac{eV}{kT}\right) < n_{p0} \tag{4–35}$$

by a similar argument. In this case, J_2^- is the electron current which flows across the junction and depletes the minority carrier level in the p-type material.

When the junction is biased in the reverse direction with a voltage greater than about 0.1 volt at 300° K, we can write

$$-eV > 4kT \quad \text{thus} \quad \exp\left(\frac{eV}{kT}\right) \ll 1$$

This corresponds to the case when J_2^- and J_2^+ carry almost all the current across the junction. From Eq. (4–5),

$$J_{\text{Di}}^+ = -eD_p \frac{dp_n}{dx}$$

where J_{Di}^+ is defined in the positive x direction, i.e., to the right in the n-type material. Hence,

$$J_{\text{Di}}^+ = -eD_p \frac{p_{n0}}{L_p} \exp\left(\frac{-x}{L_p}\right) \tag{4–36}$$

by differentiating Eq. (4–32). The negative sign in Eq. (4–36) indicates that the diffusion current flow is to the left in the n-type material.

For electrons in the p-type material, from Eq. (4–12),

$$J_{\text{Di}}^- = eD_n \frac{dn}{dx} = eD_n \frac{n_{p0}}{L_p} \exp\left(\frac{-x}{L_n}\right) \tag{4–37}$$

by differentiating Eq. (4–33). This current is defined to be in the positive x direction in the p-type material, i.e., from right to left.

At the edge of the n-type depletion region, $x = 0$, and the hole current to the left as defined in Sec. 3–5 is

$$J_2^+ = eD_p \frac{p_{n0}}{L_p} \tag{4–38}$$

At the edge of the p-type depletion region, $x = 0$, and the electron current to the left is defined as

$$J_2^- = eD_n \frac{n_{p0}}{L_n} \tag{4–39}$$

From Sec. 3–5, the total current density defined in the forward direction is

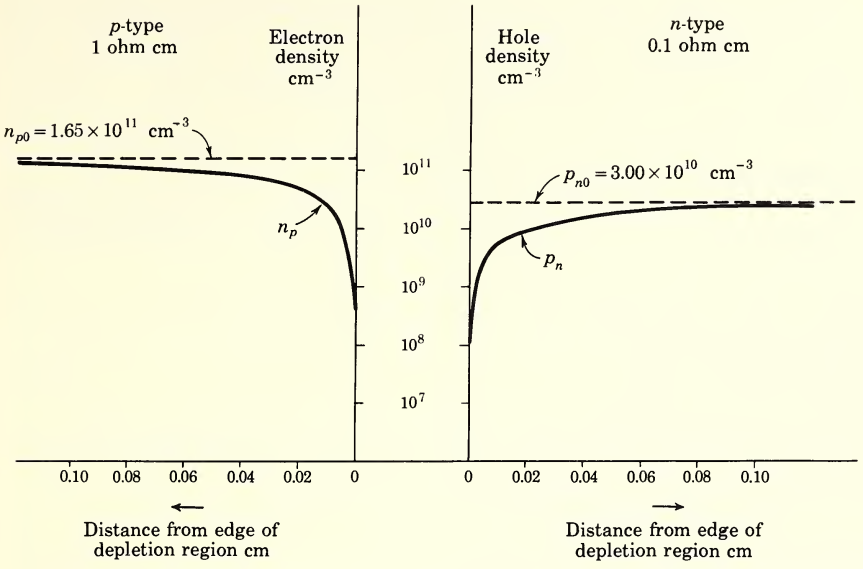
$$J = -(J_2^+ + J_2^-)$$

since we have assumed $J_1^+ = J_1^- = 0$ for a junction with a large reverse bias. Hence,

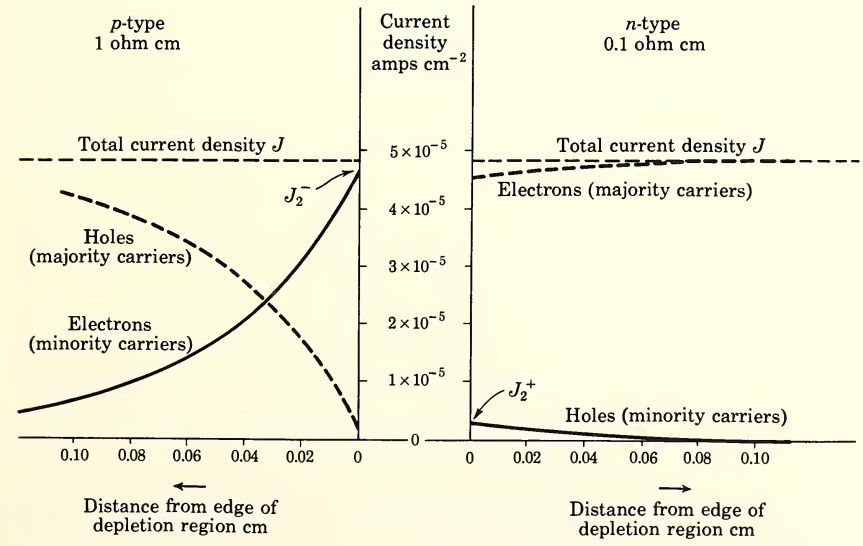
$$J = -\left(eD_p \frac{p_{n0}}{L_p} + eD_n \frac{n_{p0}}{L_n}\right) \tag{4–40}$$

Thus a reverse-biased junction will pass a current in the reverse direction given by Eq. (4–40). Since p_{n0} and n_{p0} are small compared to N_d and N_a, the reverse-biased junction current density is small.

Figure 4–5 shows the result of a computation of Eqs. (4–32), (4–33), (4–36), and (4–37) for the case of the p-n junction used as an example in Chap. 3. The minority carrier densities, which are shown in the upper set of curves, are small close to the depletion regions and increase exponentially with density in the bulk material until the thermal equilibrium value is reached. The two diffusion currents are plotted as full lines in the lower set of curves in Fig. 4–5. The diffusion current of positive holes in the n-type



(a)



(b)

FIG. 4–5. Minority carrier and current density for a reverse-biased abrupt p–n junction.

material is labeled "holes (minority carriers)" and is traveling towards the left-hand side of the diagram and constitutes a current in this direction. The diffusion flow of electrons in the p-type material is labeled "electrons (minority carriers)" and is moving to the right. Since electrons carry a negative charge, the current associated with the electron motion is to the left. In both materials, the current which is carried by the minority carriers is at its maximum at the edge of the depletion regions. Further into the bulk material, recombination takes place, and the current is mainly carried by majority carriers.

Values of the constants that have been used to draw Fig. 4–5 are given below. The voltage across the junction was -0.15 volt.

p-type germanium:

$\rho_p = 1$ ohm cm $\qquad N_a = 3.8 \times 10^{15}\ \text{cm}^{-3}$

$\mu_n = 3320\ \text{cm}^2\ \text{volt}^{-1}\ \text{sec}^{-1} \qquad n_{p0} = 1.65 \times 10^{11}\ \text{cm}^{-3}$

$D_n = 86\ \text{cm}^2\ \text{sec}^{-1} \qquad L_n = 0.05$ cm

n-type germanium:

$\rho_n = 0.1$ ohm cm $\qquad N_d = 2.1 \times 10^{16}\ \text{cm}^{-3}$

$\mu_p = 1200\ \text{cm}^2\ \text{volt}^{-1}\ \text{sec}^{-1} \qquad p_{n0} = 3.0 \times 10^{10}\ \text{cm}^{-3}$

$D_p = 31\ \text{cm}^2\ \text{sec}^{-1} \qquad L_p = 0.05$ cm

The total current density across the junction is

$$J = -\left[eD_p \frac{p_{n0}}{L_p} + eD_n \frac{n_{p0}}{L_n}\right] = -4.84 \times 10^{-5}\ \text{amps cm}^{-2}$$

This value ignores the effect of surface leakage which is present in a practical case.

4–5 Solution of the Continuity Equations for a Forward-biased Abrupt p-n Junction

The solutions of the continuity equations are limited to the small-current condition. For positive holes in the n-type material, Eq. (4–32) still applies.

$$p_n - p_{n0} = p_{n0}\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \exp\left(\frac{-x}{L_p}\right) \tag{4–32}$$

where V is now positive and $\exp(eV/kT) > 1$. Similarly, for electrons in the p-type material, Eq. (4–33) is still valid.

$$n_p - n_{p0} = n_{p0}\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \exp\left(\frac{-x}{L_n}\right) \tag{4–33}$$

where V is positive.

The diffusion current of holes at any point in the n-type material is,

$$J_{\mathrm{Di}}^{+} = -eD_p \frac{dp}{dx} = e\frac{D_p}{L_p} p_{n0}\left[\exp\left(\frac{eV}{kT}\right) - 1\right]\exp\left(-\frac{x}{L_p}\right)$$

$$= e\frac{D_p}{L_p}(p_n - p_{n0}) \tag{4–41}$$

defined in the positive x direction in the n-type material, i.e., from left to right. Similarly,

$$J_{\mathrm{Di}}^{-} = eD_n \frac{dn}{dx} = -e\frac{D_n}{L_n}(n_p - n_{p0}) \tag{4–42}$$

defined in the positive x direction in the p-type material, i.e., from right to left.

When the junction is forward biased with a voltage greater than about 0.1 volt at 300° K,

$$eV > 4kT \quad \text{and} \quad \exp\left(\frac{eV}{kT}\right) \gg 1$$

Thus the reverse saturation currents (J_2^{+} and J_2^{-}) are negligible, and the total current across the junction can be written as

$$J = J_1^{+} + J_1^{-}$$

where

$$J_1^{+} = e\frac{D_p}{L_p} p_{n0}\exp\left(\frac{eV}{kT}\right)$$

when $x = 0$ in the n-type region and J_1^{+} is defined to flow from left to right. Also,

$$J_1^{-} = e\frac{D_n}{L_n} n_{p0}\exp\left(\frac{eV}{kT}\right)$$

when $x = 0$ in the p-type region and J_1^{-} is defined to flow from left to right. Taking account of the directions of current flow, therefore,

$$J = \left(e\frac{D_p}{L_p} p_{n0} + e\frac{D_n}{L_n} n_{p0}\right)\exp\left(\frac{eV}{kT}\right) \tag{4–43}$$

The curves of charge density and current for the typical p-n junction are plotted in Fig. 4–6. The applied voltage is 0.15 volt in the forward direction. On both sides of the depletion region, the minority carrier density is increased because of the injection of minority charges from the other side of the junction. As the minority carriers diffuse away from the junction, recombination occurs and the minority carrier current density drops. For instance, in the n-type region on the right-hand side, the minority carriers (holes) recombine, and the current becomes electronic. At a distance L_n from the edge of the depletion region only 1/exp of the initial hole diffusion current is still being carried by holes.

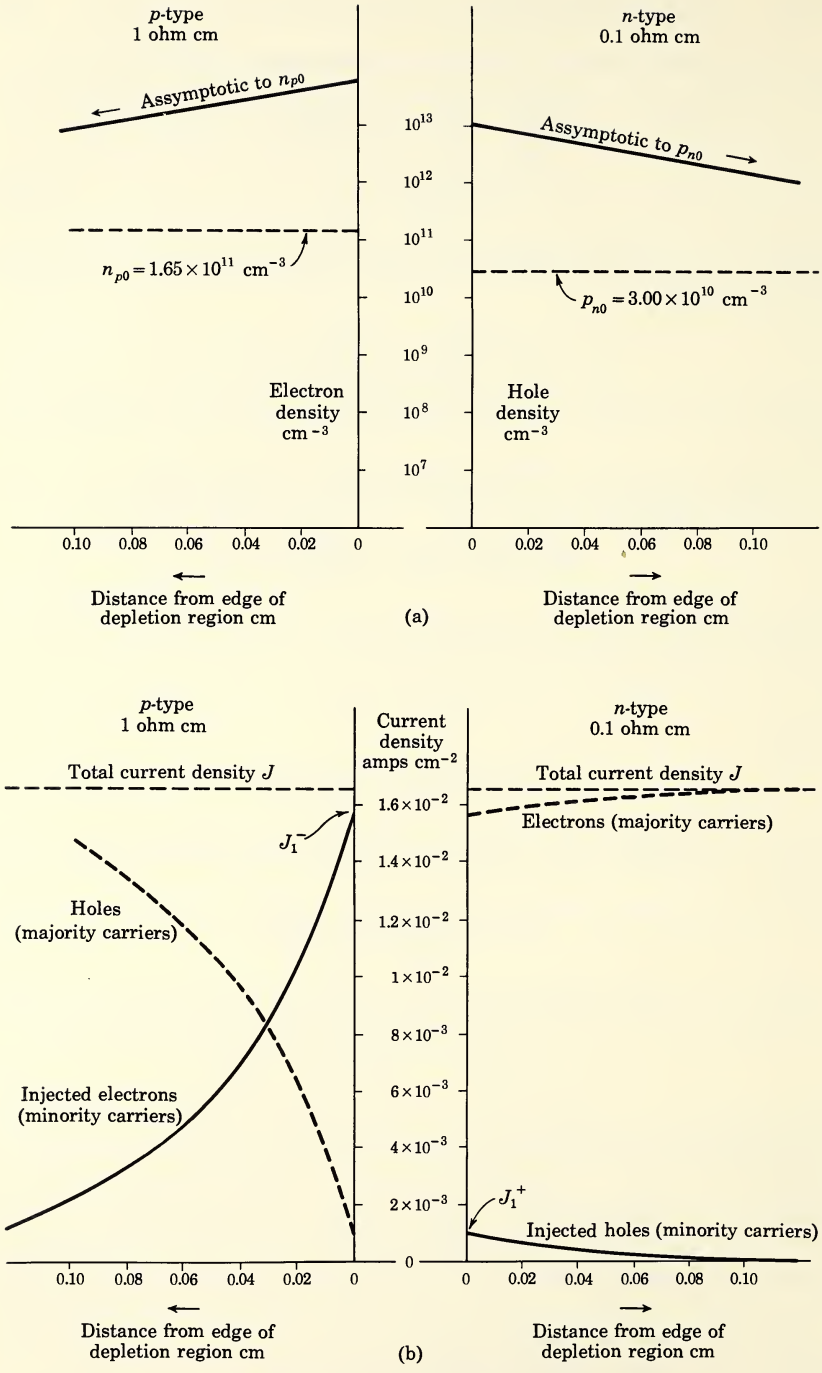


FIG. 4–6. Minority carrier and current density for a forward-biased abrupt p–n junction.

In this example, the applied voltage is assumed to be constant and so dp/dt and dn/dt are everywhere zero. Assuming isothermal and isotropic conditions throughout, the sum of the minority and majority currents is constant across all regions, and this has been indicated on the diagram. In the depletion regions, there is a scarcity of mobile charges, and recombination is very small. The hole and electron currents are assumed to be unchanged while in these regions.

4–6 A Brief Discussion of the *p-n* Junction

In the foregoing, the continuity equations have been solved for the reverse and forward bias conditions. Both cases are important in understanding the junction transistor and will be referred to in Chap. 5. At this stage, it is well to summarize briefly the physical meaning of the solutions.

When an abrupt *p-n* junction is reverse biased, *the minority carrier densities at the edge of the two depletion regions are very low*. The total reverse junction current density is also small, and it is determined by the minority carrier densities in the bulk material [Eq. (4–40)].

When the junction is forward biased, the majority of the current is carried by *injected minority carriers*. Holes will move from the *p*-type side to the *n*-type material where they become minority carriers. Electrons travel from the *n*-type material across the junction and become minority carriers in the *p*-type side. These injected currents increase the minority carrier densities close to the depletion regions, and so a diffusion of minority carriers away from the junction takes place (Fig. 4–6). As these minority carriers move away from the depletion region, recombination occurs. In the *p*-type material, the injected electrons gradually recombine with the positive holes, and the proportion of the current carried by the majority carriers (holes) increases with distance from the edge of the depletion region. At a distance $4L_n$ in the *p*-type region most of the recombination has already occurred and nearly all the current is carried by holes. In the *n*-type material, recombination also occurs with the result that the transfer of current from holes to electrons is nearly complete at a distance $4L_p$ from the edge of the *n*-type depletion region.

At this point, we can distinguish between the *p-n* junction and an ohmic contact. The sole purpose of an ohmic contact is to form a connection between the semiconductor region and the external lead so that current can be passed into, or out of, the device. Unlike the *p-n* junction, therefore, the ohmic contact should have a linear voltage versus current characteristic, and it should behave as a small ohmic resistance. Ohmic contacts are frequently made between a metal and the semiconductor region. The surface structure of the semiconductor is often deliberately damaged in some way around the contact. This increases the recombination rate in the vicinity of the contact, and, by keeping the carrier concentration near equilibrium, prevents to some extent the formation of a rectifying junction. Ohmic contacts may be formed by soldering, alloying, or other types of bonding process.

4–7 The Incremental Resistance of a Forward-biased *p-n* Junction

The total current density, J, in a forward-biased junction is given by Eq. (4–43) to be

$$J = \left(eD_p \frac{p_{n0}}{L_p} + eD_n \frac{n_{p0}}{L_n}\right) \exp\left(\frac{eV}{kT}\right) \tag{4–43}$$

For a junction of area A, the total current is

$$I = JA \tag{4–44}$$

An *incremental resistance* for the junction can be defined to be

$$r_\epsilon = \frac{dV}{dI} = \frac{1}{A}\frac{dV}{dJ} = \frac{\frac{1}{A}}{\frac{e}{kT}\left(e\frac{D_p}{L_p}p_{n0} + e\frac{D_n}{L_n}n_{p0}\right)\exp\left(\frac{eV}{kT}\right)}$$

$$= \frac{kT}{eI} \tag{4–45}$$

At room temperature, $T = 300°$ K and $kT/e = 0.0258$ volts. Hence,

$$r_\epsilon = \frac{0.0258}{I} \text{ ohms} \tag{4–46}$$

when I is specified in amps. r_ϵ is the resistance presented to small changes in the voltage across the junction. For a given junction current, the incremental resistance, from Eq. (4–46), is independent of the area of the junction.

4–8 Einstein Equation

There is an important relationship, known as the *Einstein equation*, which exists between the mobility and diffusion constant. It will be derived here by reference to the current flow in a *p-n* junction.

In the conduction band of Fig. 4–7, there are two current density components flowing across the junction, J_1^- and J_2^-. J_2^- is the current which flows down the potential hill from the *p*-type to the *n*-type region. It is, therefore, a drift current of electrons of the form,

$$J_2^- = ne\mu_n\mathscr{E} \tag{4–47}$$

where $\mathscr{E}$ is the electric field, and n is the density of electrons in the depletion regions. The current, J_1^-, is the result of diffusion of electrons from the *n*-type region, where electrons are majority carriers, to the *p*-type material, where they are in the minority. Thus,

$$J_1^- = eD_n \frac{dn}{dx}$$

We know that when the applied voltage is zero,

$$J_2^- = J_1^-$$

or

$$ne\mu_n\mathscr{E} = eD_n\frac{dn}{dx}$$

Integrating across the junction as shown in Fig. 4–7,

$$\int_a^b \mathscr{E}\,dx = \frac{D_n}{\mu_n}\int_{n_{p0}}^{N_d}\frac{dn}{n}$$

or

$$V_B = \frac{D_n}{\mu_n}\log_e\left(\frac{N_d}{n_{p0}}\right) = \frac{D_n eV_B}{\mu_n kT}$$

from Eq. (3–18). Thus,

$$\frac{\mu_n}{D_n} = \frac{e}{kT} \tag{4–48}$$

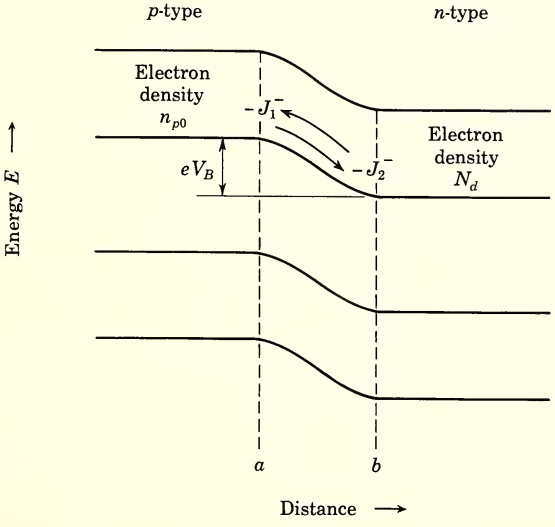


FIG. 4–7. *p–n* junction with no applied voltage showing electron densities and currents.

From a consideration of holes in the valence band, we can show that

$$\frac{\mu_p}{D_p} = \frac{e}{kT} \tag{4–49}$$

This relationship between the mobility and the diffusion constant is the Einstein equation.

BIBLIOGRAPHY

Middlebrook, R. D., *An Introduction to Junction Transistor Theory*, New York: John Wiley & Sons, Inc., 1957

Shive, J. N., *Properties, Physics and Design of Semiconductor Devices*, New York: D. Van Nostrand Company, Inc., 1959

Shockley, William, *Electrons and Holes in Semiconductors*, New York: D. Van Nostrand Company, Inc., 1950

Spangenburg, Karl R., *Fundamentals of Electron Devices*, New York: McGraw-Hill Book Company, Inc., 1957

PROBLEMS

4–1 In the derivation of the continuity equation, the effects of only three phenomena were taken into account. If the electric field in a semiconductor is related to the temperature gradient by the equation,

$$d\mathscr{E} = K\,dT$$

where $\mathscr{E}$ is the electric field, T is the temperature, and K is a constant, derive the term which must be added to the continuity equation to account for this phenomenon.

4–2 If the electric field in a semiconductor is zero and the injected electron concentration varies with an angular frequency ω, i.e.,

$$n(x, t) = n(x) \exp(j\omega t)$$

show that the continuity equation becomes

$$\frac{d^2n}{dx^2} = \left(\frac{1 + j\omega\tau_n}{L_n^2}\right)n$$

4–3 Discuss the continuity equation for electrons, identifying the terms that account for the three phenomena affecting the charge distributions and current flow in a semiconductor. How are these terms derived? What is the form of the equation if no field is applied and a steady state has been reached?

4–4 It has been stated that a very small proportion of the minority carrier current close to a semiconductor *p-n* junction is carried by drift of the minority carriers in the electric field. Using the values for the *p-n* junction given in Sec. 4–4, compare the relative magnitudes of the minority drift and diffusion currents at a distance of two diffusion lengths away from the edge of the depletion region in the *n*-type material.

4–5 A germanium *p-n* junction consists of two regions 2 mm long and having an area of 1 sq mm. The resistivity of the *n*-type region is 3 ohm cm. If the lifetime of holes in the *n*-type region is 10 microseconds and the *p*-type region is heavily doped compared to the *n*-type, calculate the approximate saturation current at 300° K.

4–6 Explain the phenomena of diffusion, drift and recombination and show how they are related in the continuity equation.

4–7 Using the data given in Sec. 4–4, determine the value of x in the n-type material at which 95 per cent of the hole minority carriers have recombined. At what distance from the edge of the depletion region do the majority and minority currents contribute equally to the total current density?

4–8 Show that Eq. (4–40) can be expressed as

$$J = -e\left(\frac{p_{n0}L_p}{J_p} + \frac{n_{p0}L_n}{J_n}\right)$$

or

$$J = -en_i^2\left(\frac{D_p}{L_pN_d} + \frac{D_n}{L_nN_a}\right)$$

4–9 Describe what is meant by "small-current" operation. Show why it is not necessary to solve the continuity equation applying to majority charges to obtain a solution for the charge density and current flow in a p-n junction operating under small-current conditions.

4–10 Show that the ratio of hole current to electron current across a semiconductor junction can be expressed as

$$\frac{J_p}{J_n} = \frac{\sigma_pL_n}{\sigma_nL_p}$$

4–11 Find the current flowing across a p-n junction of area 10^{-6} m^2 at 300° K when the applied voltage is 2 volts in the reverse direction. The resistivities of the p- and n-type regions are both 1 ohm cm. Assume $L_n = 0.1$ cm and $L_p = 0.04$ cm.

4–12 Summarize and discuss the principal electrical characteristics of a p-n junction as given in Chaps. 3 and 4, dealing in particular with (a) depletion widths, (b) incremental resistance and capacitance, (c) reverse and forward current versus voltage characteristic.

5

The Junction Transistor

In the preceding chapters the theory of electrical conduction in semiconductors has been applied only to the rectifying, abrupt *p-n* junction. In this chapter, the same theory is extended to analyze the junction transistor.

Historically, semiconducting rectifiers date back to the "crystal and cats-whisker" type of detector which was used in the early days of radio. This simple device operated as a rectifier when the cats-whisker was able to find a spot on the crystal where the impurity content was such as to create a *p-n*

junction. The development of the point-contact germanium diode prior to 1945 was the result of manufacturing a device with more control over the point contact and the semiconducting base material. In 1948, Bardeen and Brattain announced a *point-contact transistor*. This was a device having two contacts on an *n*-type germanium base, and experiment showed that such an arrangement could provide a current and power gain when operated with suitable applied potentials. This type of transistor is rarely used now because of manufacturing difficulties and the presence of a high noise level. The theory of the point-contact transistor is not completely understood at the present time.

Developments in the semiconductor field were rapid after 1945, and, by 1949, the theory of *p-n* junctions had been established. At this time, Shockley announced, and gave the theory of, a *junction transistor*, and, within a few years, this new device was commercially available. The junction transistor is composed of an *npn* or *pnp* sandwich where the junctions may be abrupt or graded. In this chapter, only transistors using abrupt junctions will be considered and the main characteristics of these devices will be discussed.

5–1 The *npn* Transistor

In the particular *p-n* junction which was used as an example in Chaps. 3 and 4, the conductivity of the *p*-type material was 1 mho cm^{-1} and that of the *n*-type material was 10 mho cm^{-1}. When this junction is forward biased with an applied voltage of 0.15 volts, the electronic current density across the junction is

$$\frac{eD_n n_{p0}}{L_n} \exp\left(\frac{eV}{kT}\right) = 1.56 \times 10^{-2} \text{ amp cm}^{-2}$$

The positive hole current density across the junction is

$$\frac{eD_p p_{n0}}{L_p} \exp\left(\frac{eV}{kT}\right) = 1.02 \times 10^{-3} \text{ amp cm}^{-2}$$

The electron current therefore carries more than 90 per cent of the charge flow across the junction. If the conductivity of the *n*-type material is increased to, say, 100 mho cm^{-1} while the *p*-type conductivity is reduced to 0.1 mho cm^{-1} nearly all of the current passing across the forward-biased junction will be carried by electrons. The energy band diagram of this junction is shown in Fig. 5–1.

When the junction is forward biased with a voltage greater than $4kT/e$, the two reverse saturation currents can be neglected, and, if we also neglect the small positive hole current in the forward direction, the total current is simply

$$J = J_1^- \tag{5–1}$$

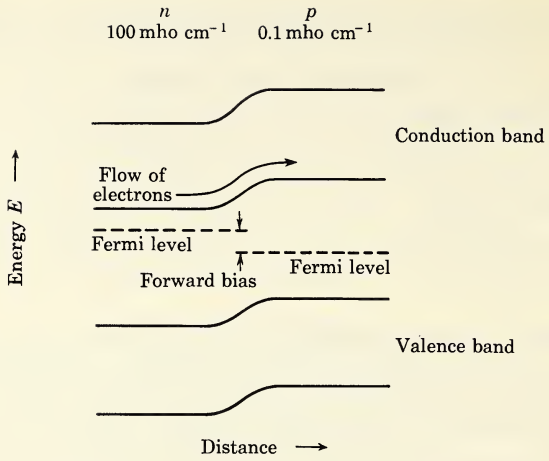


FIG. 5–1. Energy band diagram of a p–n junction where the conductivity of the n-region is much higher than that of the p-region.

The electron current J_1^- is shown in the diagram to flow from the n-type to the p-type region, Thus, the minority carrier density close to the junction in the p-type material has been increased because of electron injection.

Next consider Fig. 5–2 which represents an *npn transistor*. A second junction has been made by adding a piece of n-type material to the right-hand side of the p-type semiconductor. The width of the central p-type material in the sandwich has been made very small, and the second junction is reverse

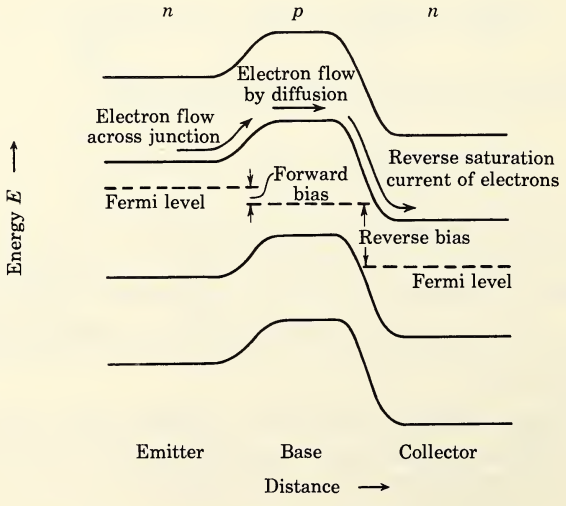


FIG. 5–2. Energy band diagram of an *npn* transistor.

biased. (The methods of manufacturing transistors are considered in Sec. 5–6.) If the width of the central region is made much smaller than the electron diffusion length, L_n, little recombination occurs before the electrons diffuse across to the second junction.

In Sec. 4–4, it was shown that the current across a reverse-biased junction is dependent upon the minority carrier density in the bulk material. In the *p*-type central region, the minority carrier density has been increased because of the injection of electrons from the left-hand junction. *The reverse saturation current of the right-hand junction is thus increased, and substantially all the injected current flows through the central region and across the second junction.*

This device is an *npn junction transistor* and the three regions are identified in Fig. 5–2 as the *emitter*, the *base*, and the *collector*. Under normal operating conditions, the emitter-to-base junction is forward biased and the base-to-collector junction is reverse biased.

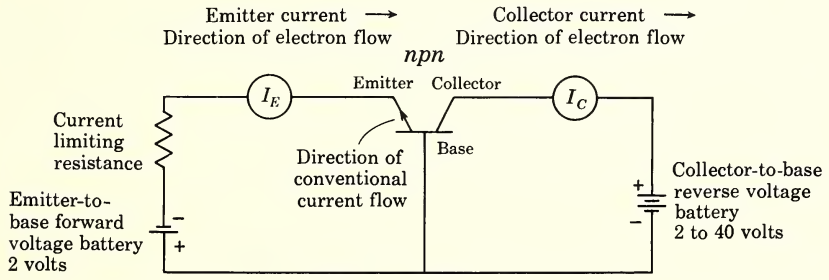


FIG. 5–3. The *npn* transistor: symbol and connections.

The flow of current across the emitter-to-base junction is known as the *emitter current*, and this current is mainly controlled by the emitter-to-base voltage. Nearly all of this current flows across the base into the collector region, and, provided the collector-to-base junction is reverse biased, the *collector current* is almost independent of the reverse applied voltage (see Fig. 5–10).

The symbol which is widely used to represent the transistor is given in Fig. 5–3 where it is seen to consist of two leads connected to a base region. This symbol was first used for the point-contact transistor and has been retained for the junction type. The emitter lead is identified by having an arrow on it; this arrow shows the direction of flow of the conventional current. For the *npn* transistor of Fig. 5–3, the electrons flow into the emitter lead, and so the direction of conventional current is out from the base.

It is not yet apparent that the transistor can be used as an amplifier. The amplifying properties of a transistor are discussed briefly in Secs. 5–4 and 5–5 and in some detail in Chaps. 6, 7, and 8. At this point, some general remarks on the operation of the transistor are in order.

It has been shown that the emitter current flows through the base and into the collector region, and it is found that the collector current is almost

independent of the collector-to-base voltage provided this junction is reverse biased. Therefore, by varying the emitter current, we will vary the collector current as well, and so control the power which will be dissipated in a load resistance placed in the collector circuit. When the collector voltage is higher than the emitter supply, the power dissipation in the collector circuit is controlled with the expenditure of a smaller power in the emitter circuit, and we have achieved a form of amplification. This is not the only, nor even the most usual, method of operating a transistor; it does indicate, however, that the device has considerable potentialities as an amplifier.

5–2 The *pnp* Transistor

Figure 5–4 shows the energy band diagrams of a *pnp transistor*. The emitter is made of high conductivity *p*-type semiconductor, and the base region is of low conductivity *n*-type material. Thus the emitter current which flows when the first junction is forward biased consists largely of positive holes which

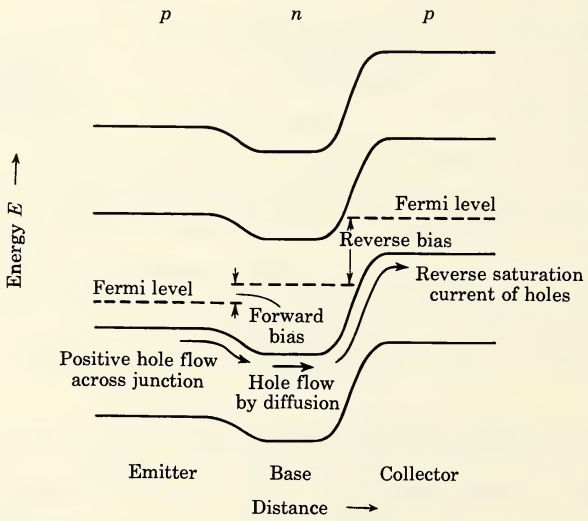


FIG. 5–4. Energy band diagrams of a *pnp* transistor.

travel through the thin base region until they reach the reverse-biased collector-to-base junction. Here the holes are collected by the final *p*-type collector region. It will be recalled that, since holes carry a positive sign, the upward direction in Fig. 5–4 represents a "downhill" or accelerating potential for holes.

The symbol used for the *pnp* transistor has the arrow on the emitter lead pointing into the base and indicating the direction of flow of conventional

current. The circuit of Fig. 5–5 is similar to that for the *npn* transistor except for the polarity of the batteries. The *pnp* transistors are used more frequently than *npn* because they are easier to manufacture (see Sec. 5–6).

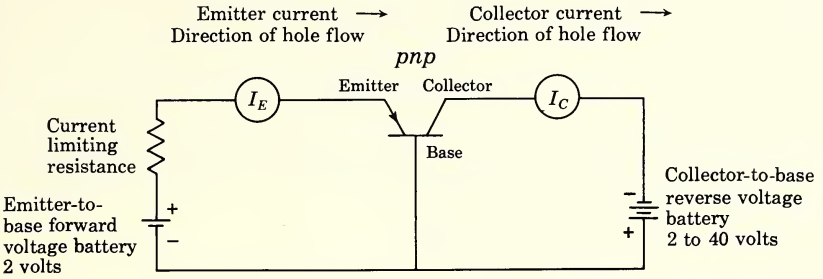


FIG. 5–5. The *pnp* transistor: symbol and connections.

5–3 Characteristics of the Current Flow across the Base Region

A. *Current flow by diffusion.* This section will summarize the principal characteristics which determine the operation of the transistor. In order to emphasize the main physical arguments several simplifying assumptions will be made. The analysis will be carried out by the application of the continuity equation to the base region of a *pnp* transistor. The analysis of an *npn* transistor follows a similar pattern.

Transistors are most frequently used as amplifiers of alternating currents in the audio and radio frequency range. Direct analysis of the transistor under a-c conditions will not be attempted in this section which is mainly concerned with direct-current operation of the device. However, the establishment of the d-c characteristics is a necessary step towards an understanding of alternating-current operation.

Consider the positive hole current which is injected into the base region of a *pnp* transistor. The continuity equation for holes in the *n*-type region states that

$$\frac{\partial p}{\partial t} = \frac{p_{n0} - p}{\tau_p} - \mu_p \frac{\partial(p\mathscr{E})}{\partial x} + D_p \frac{\partial^2 p}{\partial x^2} \tag{5–2}$$

The following simplifying assumptions will be made in the first instance:

1. Small-current, steady-state conditions apply with $\partial p/\partial t = 0$.

2. Recombination in the base region is very small since the base width is much less than L_p and so the recombination term in Eq.(5–2) can be neglected.

3. The voltage drop is confined to the two junctions so that there is no field in the base region and $\partial(p\mathscr{E})/\partial x = 0$.

Under these conditions, the continuity equation reduces to

$$D_p \frac{d^2 p}{dx^2} = 0 \tag{5–3}$$

and has a solution of the form

$$p = Ax + B \tag{5-4}$$

From Secs. 4–2 and 4–3, the hole density at the edge of an n-type depletion region in the presence of a junction voltage V is

$$P = p_{n0} \exp\left(\frac{eV}{kT}\right) \tag{5-5}$$

where p_{n0} is the thermal equilibrium hole density in the n-type base material. In the coordinate system suggested by Fig. 5–6, $x = 0$ is at the right-hand edge

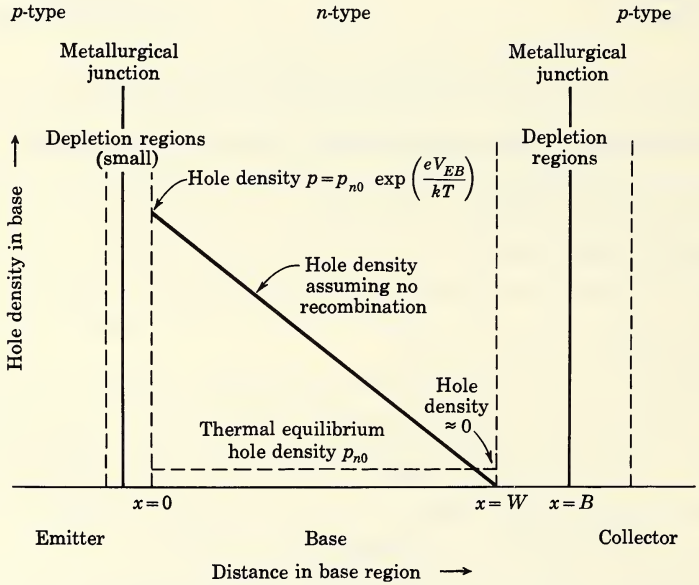


FIG. 5–6. Hole density in the base region of a *pnp* transistor assuming no recombination.

of the base-to-emitter depletion region, and, if the emitter-to-base direct voltage is V_{EB},

$$p = p_{n0} \exp\left(\frac{eV_{EB}}{kT}\right) \qquad \text{at } x = 0 \tag{5-6}$$

The edge of the collector-base depletion region is at $x = W$ in Fig. 5–6 and, if V_{CB} is the voltage across this second junction, we note that for a *pnp* transistor V_{CB} is negative and usually has a magnitude of a few volts. We can write

$$-eV_{CB} \gg kT, \text{ and so, at } x = W,$$

$$p = p_{n0} \exp\left(\frac{eV_{CB}}{kT}\right) \approx 0 \tag{5-7}$$

Substituting these two boundary values for p in Eq. (5–4),

$$p = p_{n0}\left(1 - \frac{x}{W}\right)\exp\left(\frac{eV_{EB}}{kT}\right) \tag{5–8}$$

This equation is plotted in Fig. 5–6. The three simplifying assumptions which were used at the beginning of this section, therefore, lead to the conclusion that there is a *linear reduction in hole density with distance* in the base region. The current density flowing across the base under these conditions is due entirely to *diffusion of positive holes.* The diffusion current density is

$$J = J^+ = -eD_p\frac{dp}{dx} = \frac{eD_p p_{n0}}{W}\exp\left(\frac{eV_{EB}}{kT}\right) \tag{5–9}$$

Since x does not appear in Eq. (5–9), the *diffusion current remains constant across the base region.*

B. *Emitter efficiency.* It was shown previously that by unequal doping of the p- and n-type regions of a junction the current flowing may be made to consist largely of positive holes. When this p-n junction forms the emitter-to-base junction of a *pnp* transistor, this condition must be re-examined in the light of Eq. (5–9). For positive holes moving from emitter to base, the current density is inversely proportional to the *active* base width W. For electrons moving from base to emitter, the current is given by Eq. (4–43) as

$$J_1^- = \frac{eD_n n_{p0}}{L_n}\exp\left(\frac{eV_{EB}}{kT}\right) \tag{5–10}$$

where n_{p0} is the thermal equilibrium electron density in the emitter. The ratio of the hole current to the total current flow across the junction is known as the *emitter efficiency*, γ. Since it is only the hole current which takes part in the *pnp* transistor action, γ should be very close to unity for efficient operation of the transistor.

$$\gamma = \frac{J_1^+}{J_1^+ + J_1^-} = \frac{1}{1 + \dfrac{J_1^-}{J_1^+}} = \frac{1}{1 + \dfrac{D_n n_{p0} W}{D_p p_{n0} L_n}}$$

$$\approx 1 - \frac{D_n n_{p0} W}{D_p p_{n0} L_n} = 1 - \frac{\sigma_b W}{\sigma_e L_n} \tag{5–11}$$

from Eqs. (4–48), (4–49), and (2–34), where σ_b and σ_e are the conductivities of the base and emitter regions respectively. This ratio is higher for the transistor than it is for a corresponding p–n junction. The reason for this is that under small current conditions the minority charges just at the edge of the two depletion regions are maintained at values determined by the applied voltages. The current flow of holes in the base region of the transistor, however, is greatly increased by the presence of the reverse-biased collector-to-base junction.

C. *Base width modulation.* Under small-current conditions, the hole density close to the emitter side of the base region is unaffected by the presence of a second junction at the other end of the base if V_{EB} is constant. However, the collector-to-base junction is reverse biased and acts as a sink for minority carriers in the base region. Equation (5–9) shows that the positive hole current across the base is inversely proportional to the active base width, W. In practice, the depletion region at the emitter end of the base region can be neglected since the forward bias, V_{EB} is small, but the collector-to-base de-

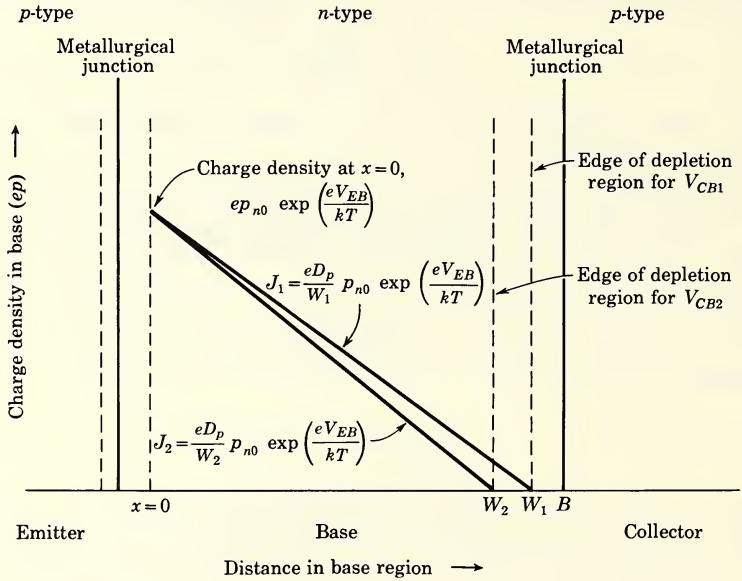


FIG. 5–7. Diagram showing the change in current density produced by a change in collector-to-base voltage, V_{CB}, when V_{EB} is a constant.

pletion width may extend a considerable distance into the base region. Writing B for the metallurgical width of the base, from Fig. 5–7, the base depletion region close to the collector is of extent B–W. From Eq. (3–13), B–W is proportional to $(V_{CB})^{\frac{1}{2}}$. Hence, W, and also the hole current through the base, are also functions of V_{CB} as shown in Fig. 5–7.

In normal operation of the transistor, the condition V_{EB} is constant is unusual, and, more commonly, the direct emitter current I_E is held constant. Under these conditions, the modulation of base width due to changes in collector-to-base voltage has only a second order effect on collector current. Small though this effect may be, it is still greater than the corresponding change in current in a simple reverse-biased p–n junction, and it is the main factor which determines the incremental collector-to-base resistance in a junction transistor. *Base width modulation* is also responsible for a small

change in the emitter-to-base voltage-current characteristic when the collector voltage changes.

D. *Recombination in the base region.* If the three assumptions listed at the beginning of this section were completely true, all the positive holes entering the base region would pass to the collector. In practice, it is found that 1 or 2 per cent of the current flows out of the transistor by the base lead. The largest component of this current is caused by recombination in the base.

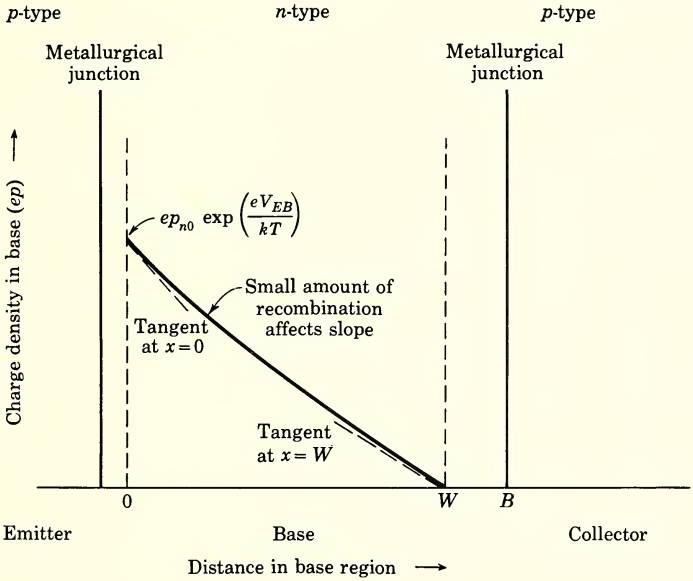


FIG. 5-8. Charge density in the base region showing the effect of recombination. [Solution of Eq. (5-12).]

Taking account of this recombination, the continuity equation can be written as

$$\frac{p - p_{n0}}{\tau_p} = D_p \frac{d^2p}{dx^2} \tag{5-12}$$

and may be solved to give a more correct analysis of the variation of hole density in the base region. This equation is not solved until Chap. 9 but Fig. 5-8 shows that the resulting charge distribution is slightly concave in shape. The hole current entering the base region is proportional to the negative slope of the curve at $x = 0$ from Eq. (5-9). From Fig. 5-8, the slope at this point is clearly more negative than the slope at $x = W$. The difference between the current entering the base region from the emitter and that leaving through the collector is the small current in the base lead.

E. *Collector-to-base cut-off current.* Another small current is flowing into the

base region which has hitherto been neglected. This is the saturation current of the collector-to-base junction which flows because the collector-to-base junction is reverse biased. This current is present even when the emitter current is zero and is known as the *cut-off current.* It is generally designated either I_{CBO} or I_{CO}. For a typical small germanium transistor at room temperature, I_{CBO} is a few microamps, and, for a silicon unit, it is usually much below one microamp. Although of small magnitude, the temperature variation of I_{CBO} is high, and, in some circuits, the cut-off current can markedly affect the total collector current (see Chap. 8).

F. *Diffusion capacitance.* When the transistor is operated in the usual manner, we have shown that the minority charge density decreases linearly with distance across the base. This charge density is established by the emitter current

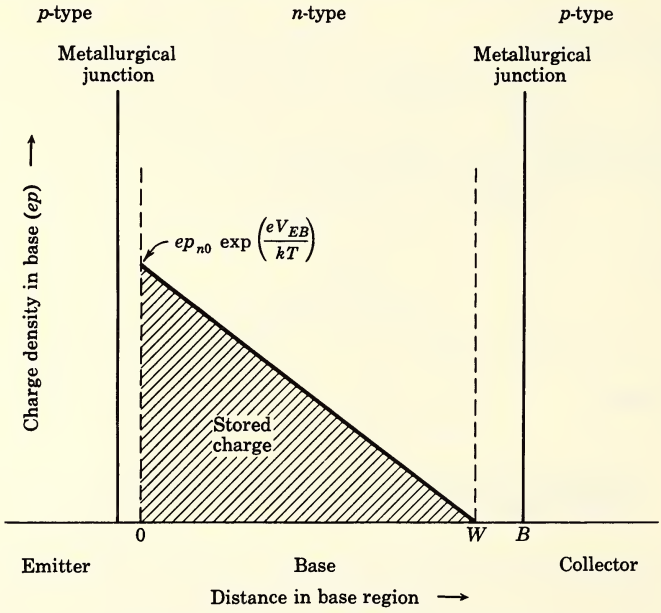


FIG. 5–9. Stored charge in the base region for the calculation of diffusion capacitance.

which flows into the base region because the emitter-to-base junction is forward biased. A change in the voltage across this junction results in a change in the minority charge in the base region, and because of this effect, the junction acts as if it had an incremental capacitance known as the *diffusion capacitance.*

From Eq. (4–31), the hole density in the base region just inside the emitter-base depletion region ($x = 0$ in Fig. 5–9) is

$$p = p_{n0} \exp\left(\frac{eV_{EB}}{kT}\right)$$

where V_{EB} is the forward bias on this junction. Ignoring recombination, the charge-versus-distance curve in the base region is a straight line as shown in Fig. 5–9. The charge stored in the base region per unit area of the junction is the shaded, triangular region in Fig. 5–9 and is given by

$$Q = \frac{ep_{n0}W}{2} \exp\left(\frac{eV_{EB}}{kT}\right) \tag{5–13}$$

where W is the active base width.

If the input voltage V_{EB} is varied, the charge stored in the base region will change. The value

$$C_D = \frac{dQ}{dV_{EB}} \tag{5–14}$$

is defined as the diffusion capacitance of the transistor per unit area of the junction. From Eqs. (5–13) and (5–14),

$$C_D = \frac{dQ}{dV_{EB}} = \frac{e^2 p_{n0}W}{2kT} \exp\left(\frac{eV_{EB}}{kT}\right)$$

But from Eq. (5–9),

$$J = \frac{eD_p}{W} p_{n0} \exp\left(\frac{eV_{EB}}{kT}\right)$$

and so

$$C_D = \frac{eW^2}{2kTD_p} J \quad \text{per unit area} \tag{5–15}$$

The incremental capacitance is given in Eq. (5–15) in terms of the current density J, and thus the total diffusion capacitance across the junction is proportional to the total emitter current, I_E.

The diffusion capacitance is in addition to the incremental capacitance of a *p–n* junction that exists because of the presence of the charge depletion regions (see Sec. 3–4). For a small transistor, the emitter-to-base diffusion capacitance is usually above 100 picofarads at $I_E = 1$ milliamp whereas the junction capacitance is typically of the order of 10 picofarads. The diffusion capacitance for the collector-to-base junction is very small since there is little alteration in the charge in the base region as the collector voltage changes. The incremental capacitance across the two junctions is a major factor in determining the upper frequency response of a transistor. This matter is discussed in some detail in Chap. 10.

5–4 The Transistor as a Device: Common Base Circuit

Figure. 5–10 shows a method of operating a *pnp* transistor. This is known as the *common base* (CB) circuit. Under normal transistor operation, the emitter-to-base junction is operated in a forward direction with a direct current I_E,

and the collector-to-base junction is biased in a reverse direction with a current flow of I_C. From the preceding analysis, I_C is slightly less than I_E and the output characteristics in Fig. 5–10 show a family of curves, each curve having a different value of I_E. The difference between I_E and I_C at any point is the base current, I_B. The output curves are nearly horizontal lines provided the collector-to-base junction is reverse biased. This indicates a high incremental output resistance of the order of 10^6 ohms. By contrast, the incremental input resistance of the transistor is of the order of a few ohms.

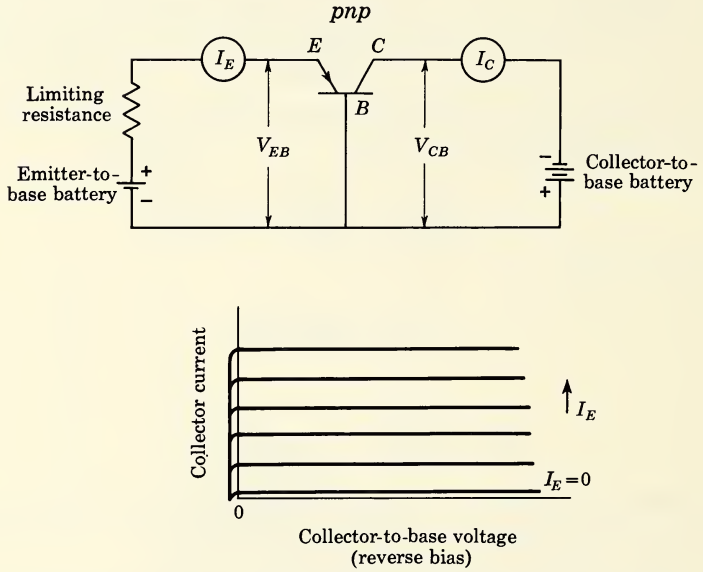


FIG. 5–10. The common base (CB) circuit and the common base output characteristics.

One of the principal components of this is the emitter-to-base junction incremental resistance (Sec. 4–7).

In most of the circuit analyses which will be carried out in the following chapters, the *current gain* of the transistor will be the factor of greatest interest (i.e., the ratio of the output and input currents). The transistor is a current-operated device, and the voltages necessary to sustain suitable currents across the junctions are of secondary importance. For example, the emitter current versus emitter-to-base voltage curve is only rarely plotted since it is so temperature dependent (see Eq. 5–9). The small emitter-to-base voltage required for forward operation of the junction is readily obtained, and interest is chiefly centered on the value of the emitter current.

Before a complete analysis of the operation of the transistor amplifier is attempted, it is of interest to show the capabilities of the device under a–c

operation. This can be done by reference to Fig. 5–11 where a small alternating-voltage generator V_e has been placed in series with the input circuit. Assuming that the alternating-voltage V_e is much less than the direct emitter-to-base voltage, a small alternating current I_e will flow in the input circuit given by

$$V_e = I_e r_i \tag{5–16}$$

where r_i is the incremental input resistance under the given conditions. If the alternating current in the output circuit is I_c, the alternating output voltage across the load is

$$V_o = I_c R_L \tag{5–17}$$

The *current gain* of the circuit is

$$A_i = \frac{I_c}{I_e} \tag{5–18}$$

and the *voltage gain* is

$$A_v = \frac{I_c R_L}{I_e r_i} \tag{5–19}$$

Equations 5–16 through 5–19 apply to small-signal a-c operation. It can be seen, however, that the ratio of the alternating currents I_c and I_e is almost the same as the ratio of the direct currents measured under the same conditions since the mechanism of charge conveyance in the base region is the same at low frequencies as at zero frequency. When the load resistance, R_L, is zero and the collector-to-base voltage is a constant, the current gain of the CB circuit is known as the α (alpha) of the transistor and has a value between 0.95 and 0.995. This is the most important transistor parameter.

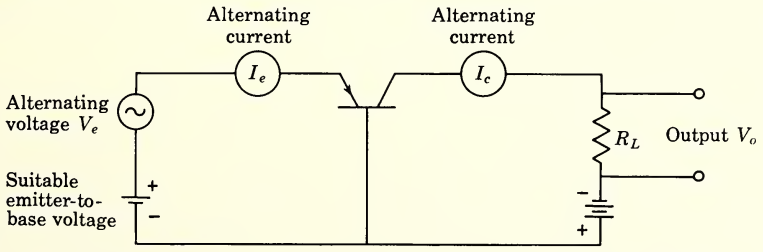


FIG. 5–11. Inclusion of a small alternating voltage in the input of a CB transistor circuit.

In the case shown in Fig. 5–11, we may approximate the condition $R_L = 0$ by $R_L < 0.1 r_o$, where r_o is the incremental output resistance, since then the voltage drop across R_L is a small proportion of the total alternating voltage in the collector circuit. Using the further approximation that $\alpha = 1$,

$$A_v = \frac{R_L}{r_i} \tag{5–20}$$

Typical values would be

$$r_o = 1 \text{ megohm}, \qquad R_L = 10{,}000 \text{ ohms}, \qquad r_i = 50 \text{ ohms}$$

giving

$$A_v = \frac{10{,}000}{50} = 200$$

From the foregoing simplified analysis, the CB transistor circuit is characterized by

1. current gain just less than unity,
2. high voltage gain,
3. low input resistance,
4. high output resistance.

5–5 The Transistor as a Device: Common Emitter Circuit

The *common emitter* (CE) circuit is the most used type of transistor amplifying stage since it provides a high current gain. To understand its principle of operation consider the three direct currents flowing in the leads of a transistor as defined by Fig. 5–12.

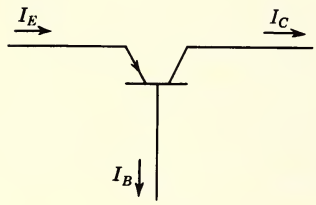


FIG. 5–12. Assigned directions of direct currents flowing in a transistor.

Neglecting any charge storage effects in the device which might occur under transient conditions,

$$I_E = I_C + I_B \tag{5–21}$$

The total collector current can be represented by

$$I_C = \alpha I_E + I_{CBO} \tag{5–22}$$

Here the collector current is written as the sum of the small cut-off current which flows even when $I_E = 0$ (see Fig. 5–10), and alpha times the emitter current I_E. The quantity α was defined for alternating-current conditions but will also apply to the direct-current case without appreciable error when the collector-to-base voltage is constant. Eliminating I_E between Eqs. (5–21) and (5–22) gives

$$I_C = \frac{\alpha}{1 - \alpha} I_B + \frac{I_{CBO}}{1 - \alpha} \tag{5–23}$$

Now assume that a small alternating current I_b is impressed upon the direct current I_B. Assuming I_{CBO} is a constant, Eq. (5–23) can be written as

$$(I_C + I_c) = \frac{\alpha}{1 - \alpha}(I_B + I_b) + \frac{I_{CBO}}{1 - \alpha} \tag{5–24}$$

Subtracting Eq. (5–23) from Eq. (5–24)

$$I_c = \frac{\alpha}{1 - \alpha} I_b \tag{5–25}$$

The current ratio I_c/I_b is the *alternating-current gain* of a transistor circuit when the base lead is used as the input and the collector lead is the output. Such an arrangement is known as a *common emitter* (CE) circuit and is shown in Fig. 5–13. This figure also shows the output characteristic curves

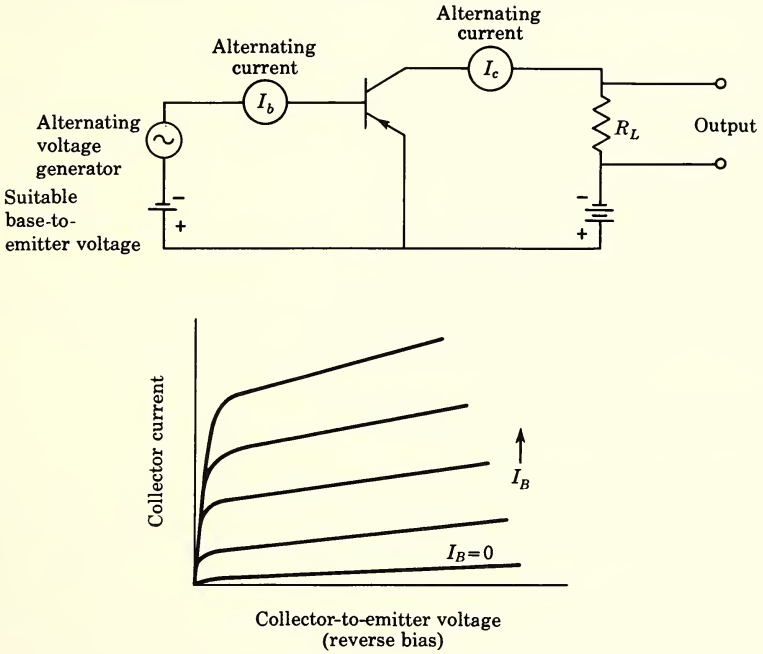


FIG. 5–13. The common emitter (CE) circuit and the common emitter output characteristics.

for this connection. By comparing the characteristic curves for the CE and the CB connection it is seen that the output incremental resistance of the CE circuit is much lower. This is a factor in the design of CE stages such as the one given in Fig. 5–13, since, to apply Eq. (5–25) directly, R_L must be much less than the incremental output resistance. Suitable values for the CE stage are

$$r_o = 20{,}000 \text{ ohms}, \qquad R_L = 1500 \text{ ohms},$$
$$r_i = 1000 \text{ ohms}, \qquad \alpha = 0.98$$

Assuming $R_L \ll r_o$, the current gain of the stage is

$$A_i = \frac{I_c}{I_b} = \frac{\alpha}{1 - \alpha} = \beta$$

The β (beta) of a transistor is the maximum current gain that can be obtained from the CE circuit and is commonly between 20 and 200. The acceptable limits of β for a particular type of transistor will be specified by the manufacturer. For the transistor mentioned above, $\beta = 49$ and so the current gain of the circuit assuming $R_L \ll r_o$ is

$$A_i = \beta = 49$$

The voltage gain through the stage is given approximately by

$$A_v = \frac{I_c R_L}{I_b r_i} = \frac{\beta R_L}{r_i} = \frac{49\,(1500)}{1000} = 73.5$$

Thus the CE transistor stage is characterized by

1. current gain fairly high (between 20 and 200),
2. voltage gain fairly high,
3. fairly low input resistance,
4. fairly high output resistance.

The CE circuit is the most used amplifier because it has a useful current gain and individual stages can be cascaded to form a multistage amplifier.

5–6 The Manufacture of Transistors

A. *Materials.* Germanium was the first material to be used in the manufacture of transistors, but in recent years increasing use has been made of silicon. Silicon has a higher forbidden gap width (1.1 volts for silicon, 0.72 volts for germanium) and so silicon devices can be operated at higher temperatures (about 175° C for silicon and 75° C for germanium). The wider band gap of silicon leads to lower leakage currents and higher impedances. Silicon has a sharper avalanche property and hence allows fabrication of voltage regulator diodes (see Sec. 3–6). In addition, silicon junctions can be made to operate at higher voltages. Finally, silicon has a remarkable natural oxide which allows easy fabrication of diffused and surface passivated devices (see Sec. 9–10).

A single crystal of germanium or silicon must be formed as a step in transistor manufacture. This crystal is grown by slowly withdrawing a "seed" crystal having the desired orientation from a crucible containing the molten semiconductor. Although chemically pure material is used, the small amount of group 3 or 5 impurity still present in the semiconductor determines its conductivity. Removal of these impurities is obtained by *zone refining*. In this technique, molten zones, obtained by localized heating, are moved along a semiconducting rod, and it is found that the impurities (being more soluble in the melt than in the solid) pass along to one end, leaving the rest of the rod with a reduced impurity content. Purity of better than 1 part in 10^{10} may be

obtained with this method. It is more difficult to produce pure single crystals from silicon than from germanium but most of the technical problems associated with the use of silicon have now been overcome.

B. *Point-contact transistors.* Many semiconducting diodes and some early transistors were made by pressing fine cats-whisker contact wires on to semiconducting wafers. Many diodes are still made in this way, but transistor production of this type has now ceased. The three major drawbacks to the use of point-contact transistors were their tendency to instability, their high noise level and the difficulty of predicting their characteristics.

C. *Grown-junction transistors.* This type of device accounts for a few per cent. of the total output of the industry. A typical germanium *npn* grown-junction transistor is cut from a bar of up to 1 in. in diameter. The single crystal bar is grown from a germanium *melt* containing arsenic to make it *n*-type. During

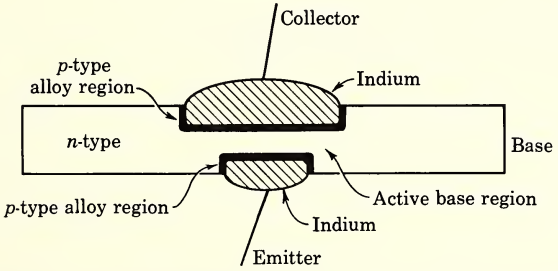


FIG. 5–14. Schematic diagram of a typical alloy-junction transistor.

the crystal growing process, gallium is added to the melt to make it *p*-type. Then, more *n*-type dope is added to make an *npn* sandwich. The bar is then sliced longitudinally into small sections about 0.025 × 0.025 × 0.125 in. long. The *p*-type base region is about 1/1000 in. thick. This device is suitable for audio-frequency operation only.

D. *Alloy-junction transistors.* Many transistors are now made using a structure similar to Fig. 5–14. Here an *n*-type semiconducting wafer of 0.08 × 0.08 × 0.004 in. thick serves as the *n*-type base. Indium is placed on the crystal wafer, and is heated to a carefully controlled temperature so that the indium dissolves (by alloying) the germanium as shown in Fig. 5–14. Upon cooling, two *p*-type regions regrow from the indium-germanium alloy and form the emitter and the collector regions. Both junctions are of *abrupt* or step-like nature. By alloying further into the *n*-type material, the thickness of the base region can be reduced. In this way, the frequency response of the device can be increased at the expense of reducing the allowed operating voltage to prevent breakdown. The manufacturing techniques become difficult when the

base thickness is reduced below a certain value and so the frequency response of the CB circuit is limited in practice to tens of megacycles per second which is adequate for many purposes. A variation of this technique involves chemically etching the base region to reduce the base thickness before alloying takes place. The device is then known as a *surface-barrier transistor*.

E. *Diffusion techniques.* This method of transistor and diode manufacture is increasingly being used in the semiconductor industry. The diffusion referred to here (doping impurity diffusion) occurs at high temperatures when there is atomic migration of group 3 or 5 material into the semiconductor. This process gives accurately controlled penetration depths, and base thicknesses of less than 5×10^{-5} in. can be produced as the difference in depth of two

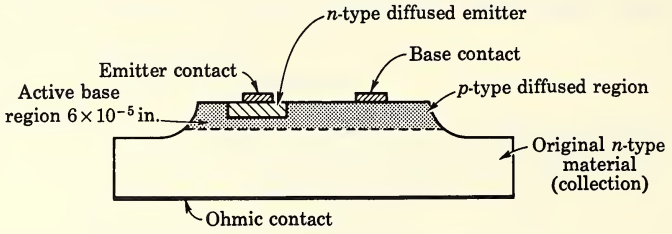


FIG. 5–15. Cross section of a mesa diffused-base transistor (simplified).

diffusion cycles. Junctions produced by this method are not abrupt but graded in some way, giving lower capacitances and often higher allowable operating voltages. Diffusants are usually in vapor form, and phosphorus (*n*-type), boron (*p*-type) and gallium (*p*-type) are commonly used. In general, hundreds of diffused devices can be made on wafers of $\frac{1}{2}$ to 1 in. diameter and then be scribed into several hundred individual devices. In this way, a highly uniform product of low cost may be obtained. A common manufacturing technique is described below.

The diffused-base "mesa" transistor is shown in simplified form in Fig. 5–15. The word "mesa" means "a flat-topped rocky hill with steeply sloping sides, common in the southwestern U.S." and describes the form of this device. The square wafer of *n*-type material shown is one of 100 to 1000 units which are made simultaneously on the same semiconducting slice. After cleaning and polishing of the top face of the slice, boron is diffused into the material to a depth of 1.6×10^{-4} in. to form a *p*-type layer. Then a second diffusion of phosphorus is made over a controlled region to a depth of 1×10^{-4} in. to give a base region of thickness 6×10^{-5} in. between the first and the second diffusions. Emitter and base metallic contacts are now evaporated on to the surface using a precision mask, and unwanted material is etched away to give the transistor its "mesa" shape. The semi-

conducting slice must still be cleaned and broken into its individual pieces before mounting and lead attachment takes place. The expense incurred in this construction is offset by the large quantity and uniformity of transistors produced from one slice of semiconductor. Diffused transistors are capable of operation above one kilomegacycle per second.

Transistors may be made using both diffusion and alloying technologies. One manufacturing process starts with a high resistance *n*-type wafer into which is diffused an *n*-type skin. An indium dot is next alloyed to one face to produce a *p-n* emitter to base junction. On the reverse face of the wafer, the *n*-type skin is removed to uncover the near-intrinsic *n*-type material below, and to this is alloyed an indium collector. This method of fabrication produces a *pnp* transistor with a graded base having the high resistance part of the base close to the collector. Thus a small collector-to-base capacitance results, and there is also a "built-in" electric field which sweeps the injected minority carriers across the base region to the collector (see Sec. 9–8). Frequency responses of greater than one kilomegacycles per second are obtainable but power dissipation is generally limited to below one watt and the cost of the device is high.

BIBLIOGRAPHY

Junction transistors

DeWitt, David, and Arthur L. Rossoff, *Transistor Electronics*, New York: McGraw-Hill Book Company, Inc., 1957

Gärtner, Wolfgang W., *Transistors: Principles, Design, and Applications*, Princeton: D. Van Nostrand Company, Inc., 1960

Greiner, R. A., *Semiconductor Devices and Applications*, New York: McGraw-Hill Book Company, Inc., 1961

Lo, Arthur W., Richard O. Endres, et al, *Transistor Electronics*, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1955

Middlebrook, R. D., *An Introduction to Junction Transistor Theory*, New York: John Wiley & Sons, Inc., 1957

Manufacture of transistors

Anderson, A. Eugene, "Transistor Technology Evolution," *Western Electric Engineer*, **3** (July 1959), 2–12, **3** (October 1959), 30–36, **4** (January 1960), 14–19

Shive, J. N., *Properties, Physics and Design of Semiconductor Devices*, New York: D. Van Nostrand Company, Inc., 1959

PROBLEMS

5–1 The *p-n* junction that was used as an example in Sec. 4–4 is now the emitter-to-base junction of a *pnp* transistor. If the base thickness is

0.002 cm, the emitter-to-base voltage is 0.2 volts in the forward direction and the collector-to-base junction is reverse biased, find the current density of holes across the base region.

5–2 Calculate the emitter efficiency for the transistor in the previous question. How can the emitter efficiency be improved?

5–3 In an alloy junction transistor, the base is a field-free region under small-current conditions. Discuss how the current passes from emitter to collector showing what conditions must be fulfilled for transistor action.

5–4 The transistor of Prob. 5–1 is operated with a collector-to-base reverse voltage of (a) 0.5 and (b) 20 volts. What are the collector currents in the two cases if V_{EB} is a constant? Assume that the base and collector conductivities are identical.

5–5 Compute the diffusion capacitance for the transistor of Prob. 5–1 when $I_E = 1$ ma.

5–6 Identify the nature of the carrier currents in each region of an alloy junction *pnp* transistor. In what manner are these currents sensitive to (a) temperature variations, (b) bias variations?

5–7 A transistor with an alpha of 0.97 is operated in a common base circuit with a resistance of 2500 ohms in the collector circuit. A small-signal a-c source is connected in series with the emitter lead. The a-c output resistance of the transistor is 1 megohm and the input resistance is 100 ohms. What are the a-c voltage and current gains for this circuit? Suggest a use for the common base circuit.

5–8 A transistor with an alpha of 0.99 is operated in a common emitter circuit. What is the maximum alternating-current gain that can be achieved? In this circuit, the direct current, I_B, is held constant as the temperature of the device changes. I_{CBO}, however, is found to double its value for a 10° C increase in temperature. If $I_B = 10$ μa (constant) and $I_{CBO} = 1$ μa at 25° C, what is the value of I_C at (a) 25° C, (b) 35° C, (c) 50° C? (Note the serious effect of changes in I_{CBO} when the CE circuit is operated with constant I_B; this effect is analyzed in Chap. 8.)

5–9 What are the effects of the conductivities of the three transistor regions on current gain, emitter efficiency and base width modulation?

5–10 It is required to cascade two transistor stages to form a multistage transistor a–c amplifier. Taking the typical stages given in Secs. 5–4 and 5–5 as a guide, design a suitable circuit for coupling (a) two common base stages, (b) two common emitter stages. Compute the current gain in both cases and discuss the results.

5–11 Describe the construction of an alloy junction and a diffused-base mesa transistor. By reference to current transistor literature, determine the capabilities of both types of transistor.

PART 2

Transistors at Low Frequencies and under d-c Conditions

6

The Transistor as a Two Port Network

In this chapter will be found a discussion of three small-signal parameter representations of the transistor. These are the z or *impedance parameters*, the y or *admittance parameters* and the h or *hybrid parameters*. Of these three, the h parameters are most commonly used, and they will be employed for circuit analysis at low frequencies in Chaps. 7 and 8. The equivalent circuits for these representations are given together with the *hybrid-π representation* which is used in Chaps. 9 and 10. The chapter opens with a general discussion of the properties of electronic amplifying devices.

6–1 General Considerations of Electronic Devices as Amplifiers

There are four requirements which must be fulfilled if an electronic device is to be a useful general-purpose amplifier. In the first place, it must have a considerable *voltage* or *current gain* and be able to control the power in its output circuit with the expenditure of relatively little power in the input circuit. Secondly, *the device must be linear* to prevent the introduction of unwanted distortion. When used as a power amplifier, a third requirement is that *this linearity must be obtainable over a substantial part of the output characteristics of the device.* The last condition is that *it must be possible to*

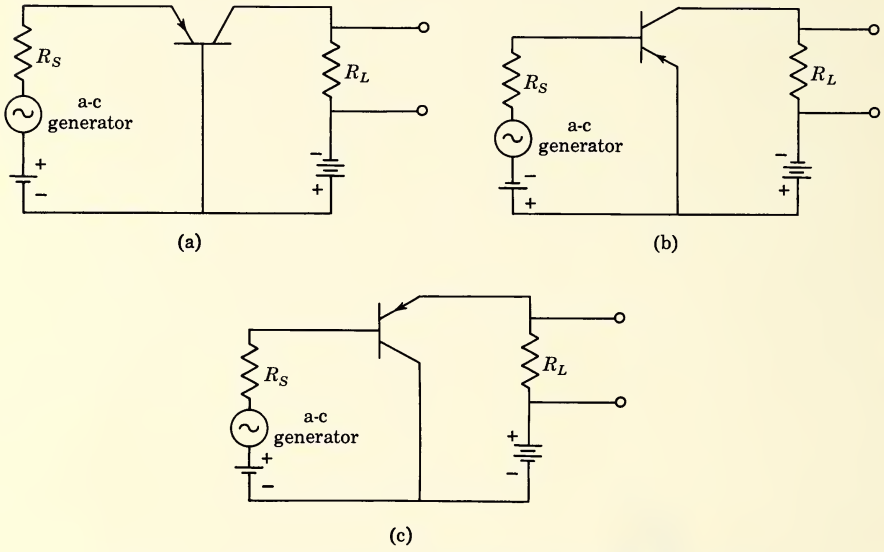


FIG. 6–1. Three transistor connections: (a) CB circuit; (b) CE circuit; (c) CC circuit.

cascade the device to form a multistage amplifier. Triode and pentode vacuum tubes meet all these conditions and any competing device must measure up to these requirements before other points of comparison are made. It will be shown that the transistor is capable of providing a high gain with a substantial amount of linearity in a cascaded amplifier, and so it can be used as a general-purpose amplifying device.

Figure 6–1 shows a transistor arranged in three possible configurations. Two of these, the CB and the CE circuits, have been briefly discussed in Secs. 5–4 and 5–5, and the third diagram shows the common collector (CC) circuit. Of these three circuits, only the CE stage can be used in multistage amplifiers. The reasons for this will be given in Chap. 7.

The transistor is mainly used as an amplifier of alternating currents, but

suitable direct-current operating conditions must also be established. This may be done by plotting the *static characteristics* of the device. Figure 6–2 shows that the transistor in any one of its three connections can be regarded as a *two port* device linking the input and output circuits. The symbols used in this figure define the *instantaneous total current and voltage* at the input and output terminals. These symbols, which conform to current practice, therefore apply when direct-current and voltage measurements are taken on the transistor to give the *input* and *output* static characteristic curves. The subscripts *e*, *b*, and *c* will be used where appropriate to denote values at the emitter, base, and collector terminals.

Consider Fig. 6–3 which shows the static input and output characteristics for the CE and CB circuits respectively. These curves are obtained from a measuring circuit of the type shown in Fig. 6–4 which applies specifically to the CE connection. As we can predict from semiconductor theory, useful operation occurs when the base-to-emitter junction has a *forward* voltage bias and when the collector-to-base junction has a *reverse*-applied voltage. We note that the reverse bias between the collector and the base is responsible for a large output resistance, whereas the forward-biased, emitter-to-base junction has a low resistance. Since the emitter current and the collector current are essentially the same, current is transferred from a low to a high resistance circuit, leading to the name *trans-resistor* or *transistor*.

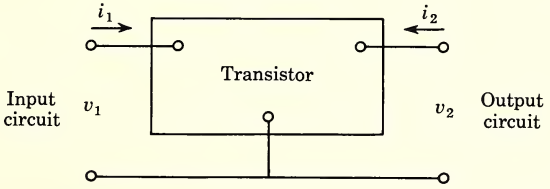


FIG. 6–2. The transistor as a two port device linking output and input circuits. Instantaneous total voltages and currents are shown.

Valuable information can be obtained from the input and output static characteristic curves of any electron device. *Voltage or current gain* can be predicted by inspection of the family of output curves. *Linearity of gain* over a wide range of input and output currents can be assessed by checking the constancy of the spacing of the output curves for equal variations in the input parameters. *Wide range of operation* is suggested by an extensive region over which the output curves have a near-constant slope. The characteristics for the CE connection of the transistor show that this method of operating the device will give a high current amplification since a small variation in i_b produces a large variation in i_c. The spacing between the output curves is approximately constant for equal variations in input current showing that the CE connection will give substantial linearity. The operating range is plainly

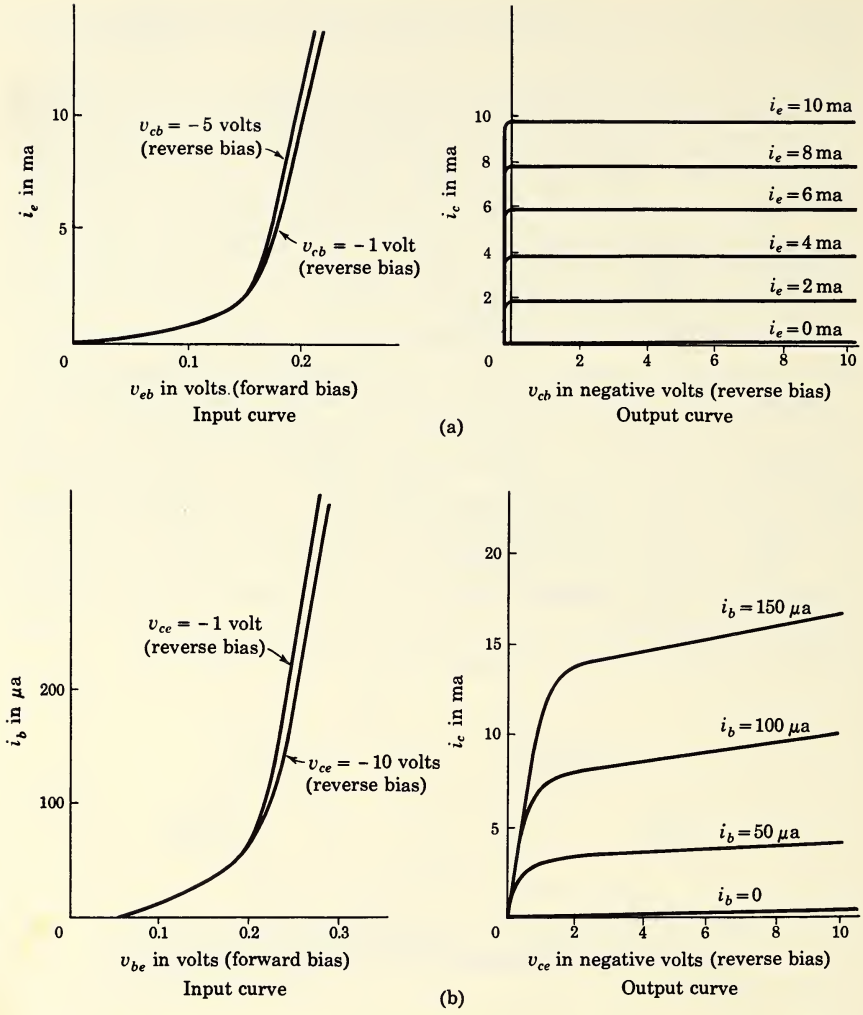


FIG. 6–3. Input and output characteristics of *pnp* germanium transistor. (a) CB connection. (b) CE connection.

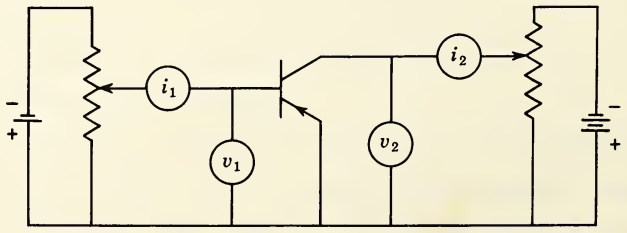


FIG. 6–4. A circuit suitable for taking static characteristic curves for the CE connection.

seen from the shape of the curves. It may be concluded, therefore, that the CE stage is a useful current amplifier. However, inspection of the base-to-emitter characteristic shows marked nonlinearity which must be taken into account in amplifier design.

The CB and the CC characteristics may also be analyzed in a similar manner. For the CB stage, it will be seen that a current gain of slightly less than unity will result, and so this stage is unsuitable for cascading in a multistage current amplifier. The CC static characteristics (which are not shown here) indicate that a high current gain can be obtained when one stage only is used, but, in Chap. 7, it is shown that this circuit, too, is never employed in multistage amplifiers.

The characteristics shown in Fig. 6–3 are the most useful ones in practice. We have reproduced the output curves with input current as a running parameter since it is known by experience that the transistor is best regarded as a *current amplifier*. The output curves drawn with input voltage as a parameter plainly show the nonlinearity which exists between input voltage and output current.

The design of single and multistage amplifiers prior to the advent of transistors had been dominated by considerations which applied to vacuum tubes. Loudspeakers, microphones, etc., had been made with impedances and power requirements suitable for triode and pentode circuits. Furthermore, matching between stages was considered solely from a voltage standpoint since the vacuum tube is primarily a voltage amplifier. However, the transistor is a *current-operated device*. It is the current gain per stage which determines the properties of a multistage transistor amplifier and the voltage gain has a much smaller significance. Consequently, since about 1950, the design of electronic equipment has been studied with regard to the use of both voltage and current amplifying devices and this has resulted in a better understanding of the processes involved.

6–2 Alternating-current Operation of the Transistor

In this, and the following chapter, only the *alternating-current* operation of the transistor will be considered. It will be assumed that the transistor is operating with a constant emitter, base, and collector current such that "small-current" conditions hold (see Sec. 4–2). The selection of a suitable operating point on the characteristic curves is discussed in Chap. 8.

There are four variables in the two port representation of a device given in Fig. 6–2. These are the two voltages $v_1$and v_2 and the two currents i_1 and i_2. The graphical relationship between these quantities is shown by a pair of the static characteristic curves of Fig. 6–3 which apply to the two port representation of the particular connection chosen. These curves determine the potentialities of the device, as described in the previous section, but for a-c analysis, it is better to consider the mathematical relationship between the

four variables. In general, an amplifier is operated in the region where the curves are straight lines so the equations should be linear.

There is some function, f_a, relating one of the variables, say v_1, with the other three. Thus

$$v_1 = f_a(v_2, i_1, i_2) \tag{6–1}$$

The quantity, v_2, could have been written in a similar general form with a different function, f_b, involving v_1, i_1, and i_2, hence

$$v_2 = f_b(v_1, i_1, i_2) \tag{6–2}$$

From Eqs. (6–1) and (6–2), we can write

$$v_1 = f_1(i_1, i_2) \tag{6–3}$$

$$v_2 = f_2(i_1, i_2) \tag{6–4}$$

where f_1 and f_2 are two functions which can be found if the functions f_a and f_b are known. Equations (6–3) and (6–4) show that i_1 and i_2 can be considered as independent variables.

By a similar analysis, the two currents can be written in terms of the two voltages to give

$$i_1 = f_3(v_1, v_2) \tag{6–5}$$

$$i_2 = \mathrm{f}_4(v_1, v_2) \tag{6–6}$$

There is no reason why one current and one voltage cannot be taken as the independent variables. The curves of Fig. 6–3 suggest the equations

$$v_1 = \mathrm{f}_5(i_1, v_2) \tag{6–7}$$

$$i_2 = \mathrm{f}_6(i_1, v_2) \tag{6–8}$$

Further equations can be written using the three remaining pairs of independent variables, but they serve no useful purpose in our analysis.

The relationships given above apply to the instantaneous total voltages and currents. For small variations in these quantities about the operating point, Eqs. (6–3) and (6–4) become

$$dv_1 = \frac{\partial v_1}{\partial i_1} di_1 + \frac{\partial v_1}{\partial i_2} di_2 \tag{6–9}$$

$$dv_2 = \frac{\partial v_2}{\partial i_1} di_1 + \frac{\partial v_2}{\partial i_2} di_2 \tag{6–10}$$

Small-signal operation applies when the variations are restricted to such a small amplitude that the partial differential terms in Eqs. (6–9) and (6–10) can be considered as constant. Inspection of the static characteristics shows how wide a range this will be. In general, this linear region should extend to at least 10 per cent of the corresponding direct operating values or the device

would not be used. Small-signal operation is usually taken to mean the conditions where all four partial differentials are constant. Linear amplification can often be obtained over a much wider range if one or more of the partial differential coefficients can be ignored.

The four partial differential terms have the dimensions of impedance and can be written

$$z_{11} = \frac{\partial v_1}{\partial i_1}, \qquad z_{12} = \frac{\partial v_1}{\partial i_2} \tag{6-11}$$

$$z_{21} = \frac{\partial v_2}{\partial i_1}, \qquad z_{22} = \frac{\partial v_2}{\partial i_2}$$

In practice, the transistor will frequently be used for the amplification of sinusoidal currents. Hence, it is convenient to replace the changes in instantaneous total voltage and current by rms a-c quantities which are considered to be impressed on the direct values. Equations (6–9) and (6–10) may now be written as

$$V_1 = z_{11}I_1 + z_{12}I_2 \tag{6-12}$$

$$V_2 = z_{21}I_1 + z_{22}I_2 \tag{6-13}$$

The four impedance terms can be defined as follows:

$z_{11} = \left.\frac{V_1}{I_1}\right|_{I_2=0}$ and is the *input impedance* when the output is open-circuited to a-c.

$z_{12} = \left.\frac{V_1}{I_2}\right|_{I_1=0}$ and is the *reverse transfer impedance* with the input open-circuited to a-c.

$z_{21} = \left.\frac{V_2}{I_1}\right|_{I_2=0}$ and is the *forward transfer impedance* with the output open-circuited to a-c.

$z_{22} = \left.\frac{V_2}{I_2}\right|_{I_1=0}$ and is the *output impedance* with the input open-circuited to a-c.

The open-circuit a-c condition assumes that the direct currents and voltages in the two port network are undisturbed. Practical methods of achieving this are discussed in Sec. 6–7.

Equations (6–12) and (6–13) are the defining equations of the *open-circuit impedance representation* of the linear two port (four terminal) network. The four impedance parameters, z_{11}, z_{12}, z_{21}, and z_{22} will be used to synthesize a circuit which represents the device within the "black box" of Fig. 6–1 as far as small-signal a-c conditions are concerned. This is the *z parameter representation.*

If the second pair of independent variables is considered, Eqs. (6–5) and (6–6) can be written in the form

$$I_1 = y_{11}V_1 + y_{12}V_2 \tag{6-14}$$

$$I_2 = y_{21}V_1 + y_{22}V_2 \tag{6-15}$$

where the y parameters are defined below

$y_{11} = \left.\frac{I_1}{V_1}\right|_{V_2=0}$ and is the *input admittance* when the output is short-circuited to a-c.

$y_{12} = \left.\frac{I_1}{V_2}\right|_{V_1=0}$ and is the *reverse transfer admittance* with the input short-circuited to a-c.

$y_{21} = \left.\frac{I_2}{V_1}\right|_{V_2=0}$ and is the *forward transfer admittance* with the output short-circuited to a-c.

$y_{22} = \left.\frac{I_2}{V_2}\right|_{V_1=0}$ and is the *output admittance* with the input short-circuited to a-c.

Here the short-circuit a-c conditions must be achieved without altering the direct currents and voltages (see Sec. 6–7). These are the *short-circuit admittance parameters* of the linear two port network and can be used to synthesize an equivalent circuit for the device under small-signal a-c conditions. This is the *y parameter representation.*

The third, and last, representation which will be given here uses the hybrid, or h, parameters. These parameters arise from considering Eqs. (6–7) and (6–8) and are defined by the two equations below.

$$V_1 = h_{11}I_1 + h_{12}V_2 \tag{6–16}$$

$$I_2 = h_{21}I_1 + h_{22}V_2 \tag{6–17}$$

where

$h_{11} = \left.\frac{V_1}{I_1}\right|_{V_2=0}$ and is the *input impedance* when the output is short-circuited to a-c.

$h_{12} = \left.\frac{V_1}{V_2}\right|_{I_1=0}$ and is the *reverse voltage amplification* factor with the input open-circuited to a-c.

$h_{21} = \left.\frac{I_2}{I_1}\right|_{V_2=0}$ and is the *forward current gain* with the output short-circuited to a-c.

$h_{22} = \left.\frac{I_2}{V_2}\right|_{I_1=0}$ and is the *output admittance* with the input open-circuited to a-c.

The parameter h_{11} has the dimensions of *impedance*, and it is equal to $1/y_{11}$. h_{22} has the dimensions of *admittance* and is equal to $1/z_{22}$. h_{12} and h_{21} are *nondimensional.* It will be shown that h_{21} is of great importance in a transistor since it is the *current gain* of the device in the short-circuit condition.

6–3 z, y, and h Parameters: General Comments

So far we have defined three sets of parameters to represent the two port network. These representations apply to small-signal operation where a linear relationship exists between the alternating currents and voltages. Since

all three representations refer to the same device, the three sets of parameters must be interrelated. At first sight, it is difficult to see why three representations of the same device are required. The reasons are partly historical and partly technical; moreover, the various manufacturers continue to specify their transistors in different ways. Historically, the z parameters were used a great deal in the early days of transistor electronics. This representation was simple and direct since an impedance was associated with each of the transistor leads. However, measurement of the z parameters is difficult (Sec. 6–7), and the representation is less useful at high frequencies. The y and h parameter representation gained in popularity with increasing usage of the transistor. In this book, the h parameter representation will be used for the

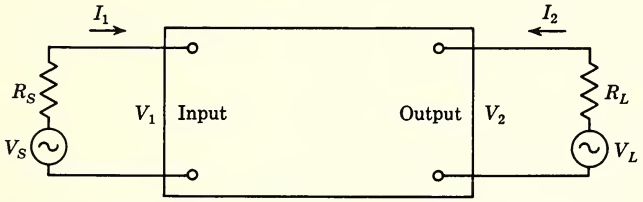


FIG. 6–5. Input and output circuits connected to the two port "black box." Rms a–c voltages and currents are shown.

analysis of low frequency operation of the device. In Part 3, where high frequency and pulse operation are considered, a modification of the y parameter representation is employed which is known as the *hybrid-π circuit*, and it will be referred to later in this chapter.

In the next three sections, equivalent circuits will be developed for the z, y, and h parameter representations respectively. In general, the justification for using these circuits will be left to the reader since this is standard two port network theory. The references given at the end of this chapter deal with this topic in detail.

The object of using any of the three sets of parameters is to provide a simple circuit representation of the transistor. Thus a transistor stage may be analyzed by replacing the device by an equivalent two port network containing resistances, capacitances, inductances, and generators. Figure 6–5 shows input and output circuits connected to the "black box." We require to know the following quantities under all circuit conditions.

Input impedance, $$Z_i = \frac{V_1}{I_1}$$

Output impedance, $$Z_0 = \frac{V_2}{I_2}$$

Current gain, $$A_i = \frac{I_2}{I_1}$$

Voltage gain, $A_v = \frac{V_2}{V_1}$

Power gain, $A_p = A_i A_v$

These terms are known as *performance quantities*, and the aim of the following circuit representation and analysis is to determine these quantities under any conditions which may exist. The voltage generator, V_L, is rarely connected in practice, but it is included in Fig. 6–5 to simplify the calculation for the output impedance.

6–4 Equivalent Circuit using z Parameters: the T Circuit

Consider the defining equations of the z parameter representation.

$$V_1 = z_{11}I_1 + z_{12}I_2 \tag{6–12}$$

$$V_2 = z_{21}I_1 + z_{22}I_2 \tag{6–13}$$

It is found that there is a simple network of impedances given in Fig. 6–6 which represents a linear two port network when $z_{12} = z_{21}$. This is known as a *passive network*. Mesh analysis of Fig. 6–6 justifies the use of the circuit.

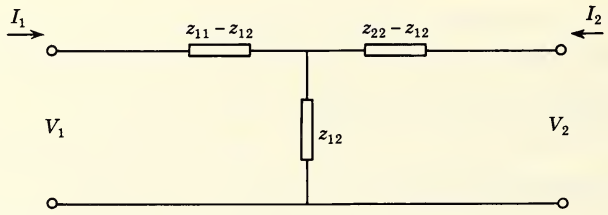


FIG. 6–6. The T circuit for a passive network.

A passive circuit can be defined as one which has no energy generating source. This circuit is not adequate for the transistor, however, since measurements show that $z_{12} \neq z_{21}$. The transistor must, therefore, be represented by an *active circuit* which contains at least one generator in addition to the impedances. Rewriting the circuit equations by adding and subtracting $z_{12}I_1$ to the right-hand side of Eq. (6–13), we have

$$V_1 = z_{11}I_1 + z_{12}I_2 \tag{6–12}$$

$$V_2 = z_{12}I_1 + z_{22}I_2 + (z_{21} - z_{12})I_1, \tag{6–18}$$

Equations (6–12) and (6–18) now relate to the network of Fig. 6–7 as can be shown by simple circuit analysis. The term $(z_{21} - z_{12})I_1$ accounts for the constant voltage generator in the circuit. When $z_{12} = z_{21}$, this generator vanishes to give the circuit of Fig. 6–6.

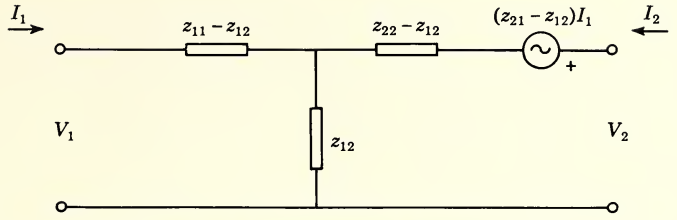


FIG. 6–7. The active T circuit.

The circuit of Fig. 6–7 may be applied directly to any of the three connections of the transistor given in Fig. 6–1. When the CB configuration is selected, the three impedances may be identified with the three leads of the transistor as shown below. At low frequencies only the resistive components of the impedances are important and so:

$$z_{11} - z_{12} = r_e, \quad \text{the emitter resistance}$$
$$z_{12} = r_b, \quad \text{the base resistance}$$
$$z_{22} - z_{12} = r_c, \quad \text{the collector resistance}$$
$$z_{21} - z_{12} = r_m, \quad \text{the transresistance}$$

Typical values for an alloy junction transistor are found to be:

$$r_e = 20 \text{ ohms}$$
$$r_b = 500 \text{ ohms}$$
$$r_c = 1 \text{ megohm}$$
$$r_m = 0.98 \text{ megohms}$$

and the circuit is given in Fig. 6–8(a). From this figure, the numerical value of the current gain is

$$\left.\frac{I_2}{I_1}\right|_{V_2 = 0} \approx \frac{r_m}{r_c}$$

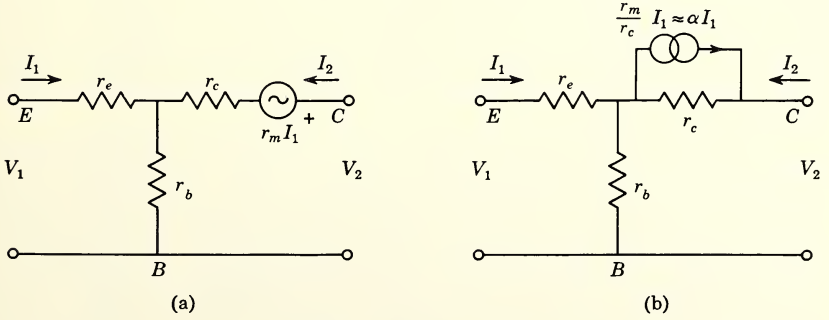


FIG. 6–8. Two common T equivalent circuits for the CB configuration.

since $r_e \ll r_b \ll r_c$. The short-circuit current gain of the device in this connection is known as α (alpha) for the transistor. For the values given above, $\alpha = 0.98$. Figure 6–8(b) shows an exactly equivalent circuit where the series (Thévenin) resistance-voltage generator combination of r_c and $r_m I_1$ has been replaced by a parallel (Norton) resistance-current generator combination formed by r_c and $I_1 r_m / r_c$. Both of these circuits will be found in transistor literature. The input current I_1 in the CB representation is the alternating emitter current, and the output current I_2 is the alternating collector current. These currents are defined as positive when flowing into their respective terminals. This is opposite to the direction of current flow which is commonly assumed in Norton and Thévenin equivalent circuits, and so the sign of the current term must always be taken into account.

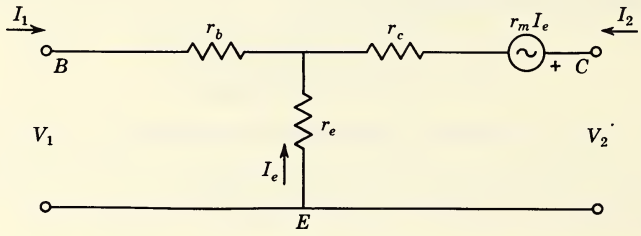


FIG. 6–9. Rearrangement of CB circuit of Fig. 6–8 (a) to make a CE circuit.

The circuits of Fig. 6–8 are seen to be a logical development of the theory given in Part 1. It was shown there, that the emitter current passes across the emitter-to-base junction, through the base region and into the collector. Either of the arrangements fits into this pattern. However, a difficulty arises since the CB circuit is rarely used as an amplifier because its current gain is less than unity. In practice, the CB circuit is retained to identify the values r_e, r_b, r_c, and r_m, but the circuit of Fig. 6–8(a) is rearranged to give the CE circuit of Fig. 6–9. Note that the conventional symbols V_1, V_2, I_1, and I_2 have been retained for the two port network, but now the input is between base and emitter and the output is taken from collector to emitter. To conform with Fig. 6–8(a), the current term controlling the constant voltage generator is now I_e not I_1.

The performance quantities defined in Sec. 6–3 can be calculated for the CB configuration by substituting either circuit of Fig. 6–8 in place of the "black box" of Fig. 6–5. For the CE circuit, Fig. 6–9 can be used. A further rearrangement of the active T circuit is required for the CC configuration. The resulting circuit analysis for any of the three cases is tedious but not difficult. The results of this analysis are given in Table 6–1. These values have been calculated at low frequencies where the impedances can be considered as resistances.

TABLE 6–1. Performance quantities at low frequences for CB, CE and CC circuits using the approximations

$$r_e \ll r_b \ll r_c \quad \text{and} \quad \beta = \alpha/(1-\alpha) \gg 1.$$

Performance quantity	*CB circuit*	*CE circuit*	*CC circuit*
Input resistance R_i	$r_e + r_b\left(1 - \dfrac{r_m}{r_c + R_L}\right)$	$r_b + \dfrac{r_e(r_c + R_L)}{r_c - r_m + R_L}$	$r_b + \dfrac{r_c(R_L + r_e)}{R_L + r_c - r_m}$
Output resistance R_o	$r_c + r_b\left(1 - \dfrac{r_m}{r_b + R_S}\right)$	$r_c - r_m + \dfrac{r_e(R_S + r_m)}{R_S + r_b}$	$r_e + \dfrac{(r_c - r_m)(R_S + r_b)}{r_c + R_S}$
Current gain A_i	$-\dfrac{\alpha}{1 + R_L/r_c}$	$\dfrac{\beta}{1 + \beta R_L/r_c}$ where $\beta = \dfrac{\alpha}{1 - \alpha}$	$\dfrac{-1}{1 - \alpha + R_L/r_c}$
Voltage gain A_v	$\dfrac{r_m R_L}{r_e(r_c + R_L) + r_b(r_c - r_m + R_L)}$	$\dfrac{-r_m R_L}{r_b(r_c - r_m + R_L) + r_e(r_c + R_L)}$	$\dfrac{r_e R_L}{r_b(r_c - r_m + R_L) + r_c(R_L + r_e)}$
Current gain for $R_L = 0$	$-\alpha$	$\beta = \dfrac{\alpha}{1 - \alpha}$	$-\dfrac{1}{1 - \alpha} \approx -\beta$

6–5 Equivalent Circuit using y Parameters: the π Circuit

Equations (6–14) and (6–15) which define the y parameters are repeated here for convenience.

$$I_1 = y_{11}V_1 + y_{12}V_2 \tag{6–14}$$

$$I_2 = y_{21}V_1 + y_{22}V_2 \tag{6–15}$$

Figure 6–10 shows the passive π circuit which applies when $y_{12} = y_{21}$.

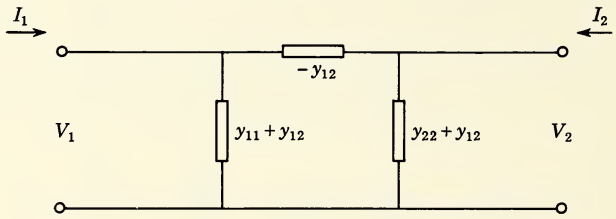


FIG. 6–10. The π circuit of a passive network.

The passive circuit needs only admittance elements in its network representation. Adding and subtracting $y_{12}V_1$ to Eq. (6–15) and rewriting Eq. (6–14), we have a pair of equations relating to the active circuit.

$$I_1 = y_{11}V_1 + y_{12}V_2 \tag{6–14}$$

$$I_2 = y_{12}V_1 + y_{22}V_2 + (y_{21} - y_{12})V_1 \tag{6–19}$$

An equivalent network of the active circuit is given in Fig. 6–11. The term $(y_{21} - y_{12})V_1$ is represented as a current generator in parallel with the output terminals, and this circuit is sometimes used to represent a transistor.

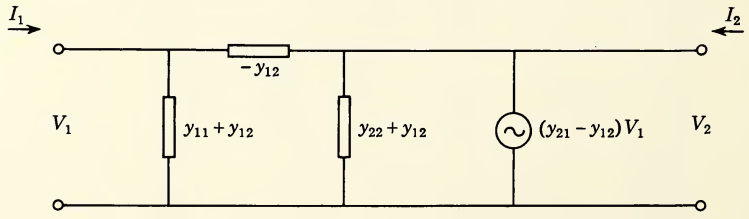


FIG. 6–11. An active π circuit.

The *hybrid-π circuit* is a useful representation of the CE connection of the transistor at high frequencies. This circuit is shown in Fig. 6–12. It is an active π circuit with an added resistance in the input lead. This resistance, which is shown between terminals B and B', is known as the *base spreading resistance* and is the ohmic resistance of the semiconducting material between the base connection and the active part of the base, B'. The hybrid-π circuit is considered in detail in Chaps. 9 and 10, where it is shown to be a good representation of the transistor at high frequencies.

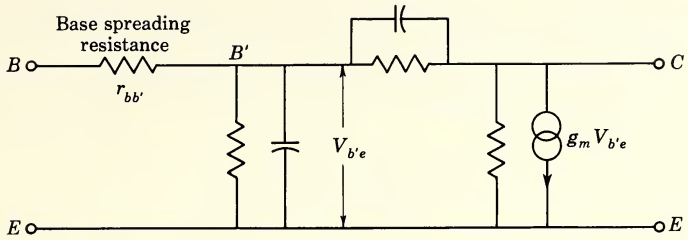


FIG. 6–12. The hybrid-π circuit.

6–6 Equivalent Circuit Using *h* Parameters: a Two Generator Representation

The third set of parameters are the ones which are most used for low frequency analysis at the present time. These are the *hybrid or h parameters* as defined by Eqs. (6–16) and (6–17).

$$V_1 = h_{11}I_1 + h_{12}V_2 \tag{6–16}$$

$$I_2 = h_{21}I_1 + h_{22}V_2 \tag{6–17}$$

where h_{11} is an impedance, h_{22} is an admittance, and h_{12} and h_{21} are nondimensional,

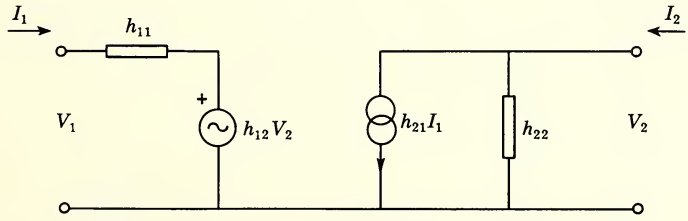


FIG. 6–13. *h* parameter equivalent circuit.

There is a two generator, two branch representation which fits the two circuit equations. This is given in Fig. 6–13. Equation (6–16) is seen to be the mesh equation of the input circuit and Eq. (6–17) is the nodal equation of the output circuit. The input contains an impedance, h_{11}, and a voltage generator, $h_{12}V_2$, in series and is thus the Thévenin equivalent of the device as seen from the input terminals. The admittance, h_{22}, and the current generator $h_{21}I_1$ comprise the Norton equivalent looking into the output terminals.

There are several reasons why *h* parameters are preferred in transistor circuitry over the other representations. As shown in the next section, they are easier to measure. They also readily fit in with the usual static characteristic curves as explained previously. h_{21}, for instance, represents the short-circuit current gain of the device which is an important quantity in transistor

amplifying stages. The Thévenin and Norton theorems are important concepts and simplify the circuit analysis. Finally, the h parameter representation leads to more simple algebraic expressions for the performance quantities. The h parameter representation will be used in the next chapter to analyze the CB, CE, and CC stages.

It is found in practice, that the two branch equivalent circuit is no disadvantage in the analysis of multistage amplifiers. It is true that by using z or y parameters a complete matrix may be set up to represent the performance quantities of such an amplifier, but such a matrix is tedious to solve and gives ample opportunities for human errors. In addition, it is shown in Chap. 7 that it is wise to design an amplifier stage by stage, starting with the output and working back to the input. For the calculation of the performance quantities of individual stages, h parameters are the most simple to use.

6–7 Low Frequency Measurements on Transistors

The d-c measuring circuit given previously in Fig. 6–4 enables the static characteristics to be drawn. To find the small-signal a-c parameters, a separate measuring circuit should be used since the z, y, or h parameters require certain specified input or output a-c conditions to be established.

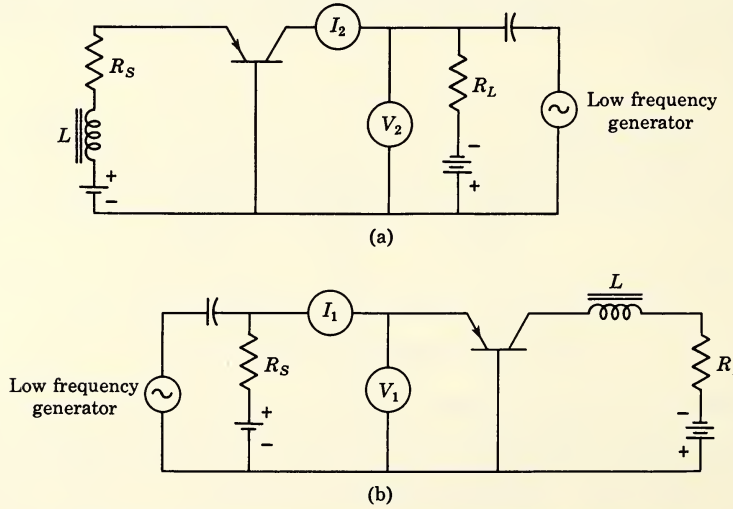


FIG. 6–14. (a) Circuit for measuring z_{22}. (b) An attempt at measuring z_{11}.

For the measurement of z parameters, the defining equations (6–12) and (6–13) show that either the input or the output circuits must be open-circuited for alternating currents. Figure 6–14(a) shows a simple circuit arranged to measure

$$z_{22} = \left.\frac{V_2}{I_2}\right|_{I_1 = 0}$$

in the CB configuration. The input is open-circuited by means of an inductance L. The reactance of the coil must be much greater than the input resistance of the transistor which, in practice, is only a few ohms. Even at low frequencies, this can readily be achieved. Figure 6–14(b) shows an attempted measurement of

$$z_{11} = \left.\frac{V_1}{I_1}\right|_{I_2 = 0}$$

The output resistance of a transistor is usually about one megohm, and thus the impedance in series with the output should be 100 megohms or greater. A resistance of this magnitude cannot be employed because it would disturb the d-c conditions. Using a choke in the position shown in Fig. 6–14(b) is not satisfactory since the impedance of the choke cannot be made high enough to approximate to the open-circuit condition. For alloy-junction transistors, frequencies of less than 1000 cps must be used to eliminate reactive effects in the device, and so the open-circuit output condition cannot be achieved.

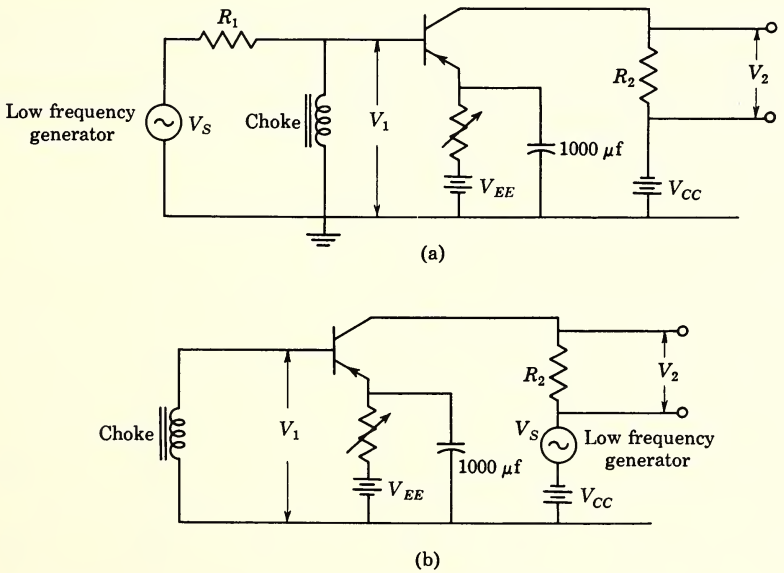


FIG. 6–15. (a) Circuit for measuring h_{ie} and h_{fe}. (b) Circuit for measuring h_{oe} and h_{re}.

Measurement of the y parameters requires a short-circuit a-c condition. In the output circuit, this can be obtained by shunting a large capacitor across the output terminals. In the input circuit, it is necessary to produce an a-c short-circuit having an impedance much less than the input resistance of the transistor. This condition is difficult to achieve since the input resistance is usually only a few ohms.

Inspection of the definitions of the h parameters given in Eqs. (6–16) and (6–17) shows that measurement of all four h parameters is relatively easy. Determination of h_{11} and h_{21} can be carried out when the output circuit is short-circuited to a-c and h_{12} and h_{22} can be found when the input is open-circuited. These two conditions are easy to reproduce, and so the h parameter measurements are the ones most frequently taken. In many cases it is desirable to measure the CE hybrid parameters directly. This can be done by using the circuits of Fig. 6–15. Figure 6–15(a) may be used to determine h_{ie} and h_{fe}. Assuming R_1 is of the order of one megohm and R_2 is a few ohms,

$$h_{ie} = \frac{R_1 V_1}{V_S} \quad \text{and} \quad h_{fe} = \frac{V_2 R_1}{V_S R_2}$$

In Fig. 6–15(b), the input circuit is open-circuited as far as a-c is concerned, and

$$h_{oe} = \frac{V_2}{R_2 V_S} \quad \text{and} \quad h_{re} = \frac{V_1}{V_S}$$

The relative positions of resistance R_2, battery V_{CC}, and generator V_S, which are in series in the output circuit is determined by the type of supplies and the voltmeter available. It is often necessary to ground one side of the voltmeter terminals to avoid stray pickup. This can be accomplished by using a shielded

TABLE 6–2. Relationship between the parameters.

Matrix ↓	*In terms of* y		z		h	
y	y_{11}	y_{12}	$\frac{z_{22}}{\Delta^z}$	$\frac{-z_{12}}{\Delta^z}$	$\frac{1}{h_{11}}$	$\frac{-h_{12}}{h_{11}}$
	y_{21}	y_{22}	$\frac{-z_{21}}{\Delta^z}$	$\frac{z_{11}}{\Delta^z}$	$\frac{h_{21}}{h_{11}}$	$\frac{\Delta^h}{h_{11}}$
z	$\frac{y_{22}}{\Delta^y}$	$\frac{-y_{12}}{\Delta^y}$	z_{11}	z_{12}	$\frac{\Delta^h}{h_{22}}$	$\frac{h_{12}}{h_{22}}$
	$\frac{-y_{21}}{\Delta^y}$	$\frac{y_{11}}{\Delta^y}$	z_{21}	z_{22}	$\frac{-h_{21}}{h_{22}}$	$\frac{1}{h_{22}}$
h	$\frac{1}{y_{11}}$	$\frac{-y_{12}}{y_{11}}$	$\frac{\Delta^z}{z_{22}}$	$\frac{z_{12}}{z_{22}}$	h_{11}	h_{12}
	$\frac{y_{21}}{y_{11}}$	$\frac{\Delta^y}{y_{11}}$	$\frac{-z_{21}}{z_{22}}$	$\frac{1}{z_{22}}$	h_{21}	h_{22}

$$\Delta^y = y_{11}y_{22} - y_{12}y_{21}$$
$$\Delta^z = z_{11}z_{22} - z_{12}z_{21}$$
$$\Delta^h = h_{11}h_{22} - h_{12}h_{21}$$

1:1 transformer with its primary connected across R_2, and its secondary connected to the voltmeter. If an oscilloscope is used as the measuring instrument, it may be advisable to ground one side of R_2 and use an audio generator and a power supply which can be operated above ground potential.

Table 6–2 shows the relationship between the sets of parameters and can be used to convert between any one set of parameters and the others.

BIBLIOGRAPHY

Hurley, Richard B., *Junction Transistor Electronics*, New York: John Wiley & Sons, Inc., 1958

Riddle, Robert L. and Marlin P. Ristenbatt, *Transistor Physics and Circuits*, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1958

Skilling, Hugh Hildreth, *Electrical Engineering Circuits*, New York: John Wiley & Sons, Inc., 1957

PROBLEMS

6–1 Verify the performance quantities in Table 6–1 for the common emitter stage using the T equivalent circuit.

6–2 A CE stage has the following parameters: $r_b = 750$ ohms, $r_e = 13$ ohms, $r_c = 1.5$ megohms, $r_m = 1.47$ megohms. If the resistance, R_S, of the input generator is 1000 ohms and $R_L = 5000$ ohms, draw an equivalent T circuit and find the current gain. Compare the result obtained with the value found by substitution in the appropriate equation in Table 6–1.

ans. 42.0, 42.6

6–3 What are the characteristics of a useful general purpose amplifier? Explain what is meant by "small-signal" conditions. What are the limitations placed on z_{11}, z_{12}, z_{21}, and z_{22} if the device is to be a linear current amplifier over a wide range?

6–4 By circuit analysis, verify that Eqs. (6–12) and (6–18) apply to Fig. 6–7 and that Eqs. (6–14) and (6–19) apply to Fig. 6–11.

6–5 Derive the values of h_{11}, h_{12}, h_{21}, and h_{22} in terms of the z parameters and check your results from Table 6–2.

6–6 Which internal equivalent impedances affect transistor operation at high frequencies? Identify these impedances in Fig. 6–12, and justify your answer.

6–7 Why is it difficult to measure y and z parameters? Explain why it is easier to measure the hybrid parameters.

6–8 Describe experiments to measure the h parameters giving the values of all circuit components used.

7

Low Frequency *h* Parameter Representation

In this chapter, the *h* parameters are used in an analysis of the single-stage transistor amplifier. The performance quantities are calculated for the general representation, and then the CE, CB, and CC stages are compared. The discussion is limited to low frequencies, where the reactive effects of the transistor can be neglected. Curves of the variation in the *h* parameters of a typical transistor with emitter current and collector voltage are given.

7–1 The Small-signal *h* Parameters

In Sec. 6–2, the *h* parameters were defined by the two equations:

$$V_1 = h_{11}I_1 + h_{12}V_2 \tag{7–1}$$

$$I_2 = h_{21}I_1 + h_{22}V_2 \tag{7–2}$$

where h_{11} = input impedance,

h_{12} = reverse voltage gain,

h_{21} = forward current gain,

h_{22} = output admittance.

h_{11} and h_{21} are measured under a-c short-circuit output conditions. h_{12} and h_{22} are measured under a-c open-circuit input conditions. The general small-signal two generator representation is shown in Fig. 7–1.

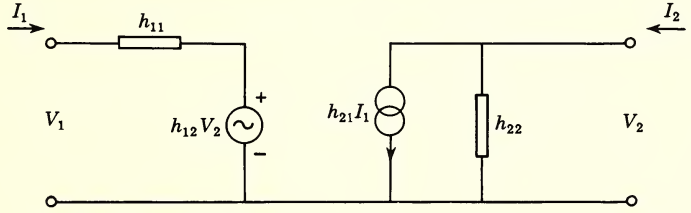


FIG. 7–1. The small-signal *h* parameter equivalent circuit.

Equations (7–1) and (7–2) and the small-signal equivalent circuit apply to all three connections, CB, CE, and CC. However, the values of the parameters will be different for the three cases. To distinguish among them, the lettered subscripts shown in Table 7–1 are used. The first letter of the subscript describes the term (input, reverse, forward, or output) and the second letter denotes the circuit connection (base, emitter, or collector). Figure 7–2 shows the *h* parameter equivalent circuit for the three connections.

In this chapter, we will use the numbered parameters for the analysis of the general equivalent circuit, and then substitute the appropriate lettered parameter for the actual circuit under consideration.

TABLE 7–1. Numbered and lettered *h* parameters.

Term	*Numbered parameter*	*CB parameter*	*CE parameter*	*CC parameter*
Input resistance	h_{11}	h_{ib}	h_{ie}	h_{ic}
Reverse voltage gain	h_{12}	h_{rb}	h_{re}	h_{rc}
Forward current gain	h_{21}	h_{fb}	h_{fe}	h_{fc}
Output admittance	h_{22}	h_{ob}	h_{oe}	h_{oc}

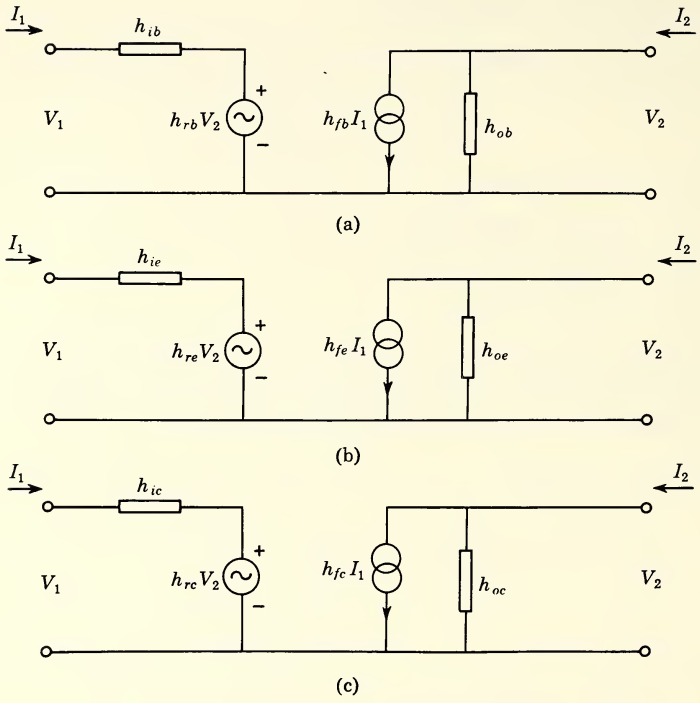


FIG. 7–2. *h* parameter equivalent circuits: (a) CB circuit; (b) CE circuit; (c) CC circuit.

7–2 Analysis of the General *h* Parameter Equivalent Circuit

Figure 7–3 shows the *h* parameter equivalent circuit of the transistor with applied voltage sources and series resistances in the input and output circuits (see also Fig. 6–5).

The five performance quantities below will be calculated at low frequencies.

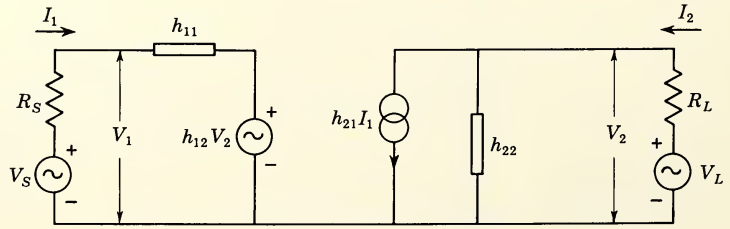


FIG. 7–3. General equivalent circuit of a transistor amplifying stage.

Input resistance, $R_1 = \dfrac{V_1}{I_1}$ Current gain, $A_i = \dfrac{I_2}{I_1}$

Output resistance, $R_o = \dfrac{V_2}{I_2}$ Voltage gain, $A_v = \dfrac{V_2}{V_1}$

Power gain, $A_p = A_i A_v$

For the input circuit of Fig. 7–3

$$V_1 = h_{11}I_1 + h_{12}V_2 \tag{7-1}$$

and for the output circuit

$$V_2 h_{22} + h_{21}I_1 = I_2 = \frac{V_L - V_2}{R_L}$$

or

$$V_L = h_{21}R_L I_1 + V_2(1 + h_{22}R_L) \tag{7-3}$$

A. *Input resistance.* Solving Eqs. (7–1) and (7–3) for I_1 gives

$$I_1 = \frac{1}{D}\begin{vmatrix} V_1 & h_{12} \\ V_L & 1 + h_{22}R_L \end{vmatrix}$$

where D $= \begin{vmatrix} h_{11} & h_{12} \\ h_{21}R_L & 1 + h_{22}R_L \end{vmatrix}$

$$= h_{11}(1 + h_{22}R_L) - h_{12}h_{21}R_L$$

$$= \Delta^h R_L + h_{11}$$

if $\Delta^h = h_{11}h_{22} - h_{12}h_{21}$

When $V_L = 0$,

$$I_1 = \frac{V_1(1 + h_{22}R_L)}{\Delta^h R_L + h_{11}}$$

and so

$$R_i = \frac{V_1}{I_1} = \frac{\Delta^h R_L + h_{11}}{1 + h_{22}R_L} \tag{7-4}$$

B. *Voltage gain.* When $V_L = 0$, from Eqs. (7–1) and (7–3)

$$V_2 = \frac{1}{D}\begin{vmatrix} h_{11} & V_1 \\ h_{21}R_L & 0 \end{vmatrix} = -\frac{h_{21}V_1R_L}{\Delta^h R_L + h_{11}}$$

Thus,

$$A_v = \frac{V_2}{V_1} = -\frac{h_{21}R_L}{\Delta^h R_L + h_{11}} \tag{7-5}$$

C. *Current gain.* In the output circuit of Fig. 7–3 with $V_L = 0$,

$$V_2 = -I_2 R_L = \frac{I_2 - h_{21} I_1}{h_{22}}$$

since the current $I_2 - h_{21} I_1$ is flowing through the admittance h_{22}. Hence,

$$A_i = \frac{I_2}{I_1} = \frac{h_{21}}{1 + h_{22} R_L} \tag{7–6}$$

D. *Output resistance.* In the input circuit of Fig. 7–3, when $V_S = 0$, $V_1 = -I_1 R_S$, and so Eq. (7–1) can be rewritten as

$$0 = I_1(h_{11} + R_S) + h_{12} V_2 \tag{7–7}$$

The output equation (7–2) for the two port network is unchanged and is

$$I_2 = h_{21} I_1 + h_{22} V_2 \tag{7–2}$$

Hence,

$$V_2 = \frac{\begin{vmatrix} h_{11} + R_S & 0 \\ h_{21} & I_2 \end{vmatrix}}{\begin{vmatrix} h_{11} + R_S & h_{12} \\ h_{21} & h_{22} \end{vmatrix}} = \frac{I_2(h_{11} + R_S)}{(h_{11} + R_S)h_{22} - h_{12} h_{21}}$$

giving

$$R_o = \frac{V_2}{I_2} = \frac{h_{11} + R_S}{\Delta^h + h_{22} R_S} \tag{7–8}$$

E. *Power gain.*

$$A_p = A_v A_i = \frac{-h_{21}^2 R_L}{(\Delta^h R_L + h_{11})(1 + h_{22} R_L)}$$

These results are summarized in Table 7–2.

TABLE 7–2. Performance quantities for the *h* parameter equivalent circuit.

Performance quantity	*Symbol*	*Value*
Input resistance	R_i	$\frac{\Delta^h R_L + h_{11}}{1 + h_{22} R_L}$
Output resistance	R_o	$\frac{h_{11} + R_S}{\Delta^h + h_{22} R_S}$
Current gain	A_i	$\frac{h_{21}}{1 + h_{22} R_L}$
Voltage gain	A_v	$\frac{-h_{21} R_L}{\Delta^h R_L + h_{11}}$
Power gain	A_p	$\frac{-h_{21}^2 R_L}{(\Delta^h R_L + h_{11})(1 + h_{22} R_L)}$

$$\Delta^h = h_{11} h_{22} - h_{12} h_{21}$$

Before the values of these quantities are determined for the three specific connections, the relationship of the parameters of the three circuits will be discussed.

7–3 Relationship of the Three Sets of *h* Parameters

It will be shown later that the CE circuit is the most used transistor amplifier. Therefore the CE parameters, h_{ie}, h_{fe}, h_{re}, and h_{oe} are of most interest. Transistor manufacturers vary widely as to which parameters they specify.

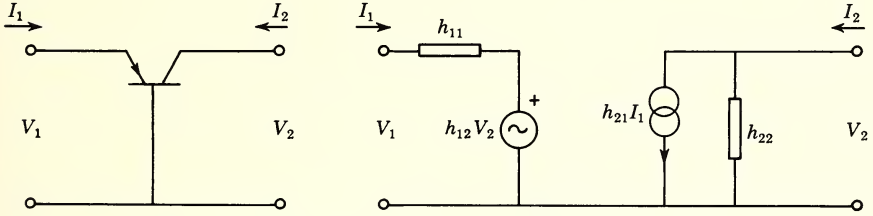


FIG. 7–4. CB equivalent circuit using numbered parameters.

The current gain, h_{fe}, will always be given since it is of great importance, but the other three may sometimes be omitted. A few years ago only CB parameters were specified, and it was necessary to convert from one set of parameters to the other. Figures 7–4 and 7–5 show how this conversion may be

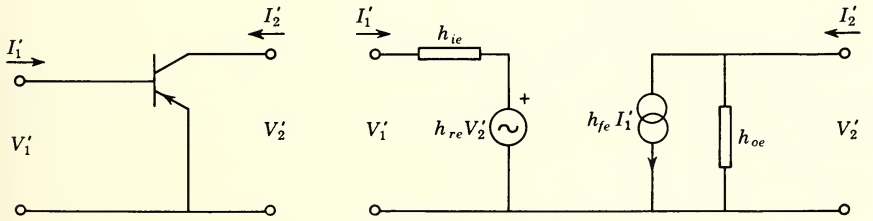


FIG. 7–5. CE equivalent circuit using lettered parameters.

effected. The CB circuit using numbered parameters is shown in Fig. 7–4. (Whenever numbered parameters are shown without further description, they refer to the CB connection.) Figure 7–5 gives the similar equivalent circuit for the CE connection where lettered parameters are used. The CB circuit may be arranged to give a CE connection as in Fig. 7–6. This last figure must correspond exactly with Fig. 7–5 where

$$V_2 = V_2' - V_1'$$

and I_1 is the current flowing upwards through h_{11}, that is,

$$I_1 = -(I_1' + I_2')$$

By means of this equivalence, it can be shown that

$$h_{ie} = \frac{h_{11}}{\Delta^h + h_{21} - h_{12} + 1}$$

$$h_{fe} = \frac{\Delta^h + h_{21}}{\Delta^h + 1 + h_{21} - h_{12}}$$

$$h_{re} = \frac{\Delta^h - h_{12}}{\Delta^h + h_{21} - h_{12} + 1}$$

$$h_{oe} = \frac{h_{22}}{\Delta^h + h_{21} - h_{12} + 1}$$

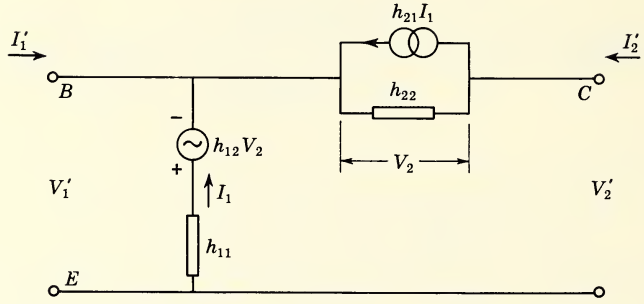


FIG. 7–6. The CB numbered h parameter equivalent circuit rearranged as a CE connection.

For a typical transistor,

$$\Delta^h - h_{12} \ll 1 + h_{21} \quad \text{and} \quad h_{21} \gg \Delta^h$$

and these equations may be simplified to the values given in Table 7–3. The corresponding equations relating the CC h parameters to those of the CB connection are also given in this table.

TABLE 7–3. h parameter equivalents of CE and CC circuits in terms of CB parameter (approximate values).

CB parameter numbered	*CB parameter* lettered	*CE parameter*	*CC parameter*
h_{11}	h_{ib}	$h_{ie} = \frac{h_{11}}{1 + h_{21}}$	$h_{ic} = \frac{h_{11}}{1 + h_{21}}$
h_{12}	h_{rb}	$h_{re} = \frac{\Delta^h - h_{12}}{1 + h_{21}}$	$h_{rc} = 1$
h_{21}	h_{fb}	$h_{fe} = \frac{-h_{21}}{1 + h_{21}}$	$h_{fc} = \frac{-1}{1 + h_{21}}$
h_{22}	h_{ob}	$h_{oe} = \frac{h_{22}}{1 + h_{21}}$	$h_{oc} = \frac{\cdot h_{22}}{1 + h_{21}}$

7–4 Comparison of the Three Configurations

Three of the performance quantities which determine the properties of the amplifying stage have been given as a log-log plot in Figs. 7–7, 7–8, and 7–9. Curves are shown for all three connections using a typical transistor having h parameters values as given in Table 7–4.

TABLE 7–4. Typical h parameter values.

Parameter	*CB value*	*CE value*	*CC value*
h_{11}	$h_{ib} = 40$ ohms	$h_{ie} = 2000$ ohms	$h_{ic} = 2000$ ohms
h_{12}	$h_{rb} = 4 \times 10^{-4}$	$h_{re} = 16 \times 10^{-4}$	$h_{rc} \approx 1$
h_{21}	$h_{fb} = -0.98$	$h_{fe} = 49$	$h_{fc} = -50$
h_{22}	$h_{ob} = 10^{-6}$ mhos	$h_{oe} = 5 \times 10^{-5}$ mhos	$h_{oc} = 5 \times 10^{-5}$ mhos
Δ^h	$\Delta^{hb} = 432 \times 10^{-6}$	$\Delta^{he} = 0.0218$	$\Delta^{hc} = 50.1$

A. *Common base circuit.* In Fig. 7–7, the current gain A_i is plotted as a function of the load resistance R_L over the wide range, $R_L = 100$ ohms to

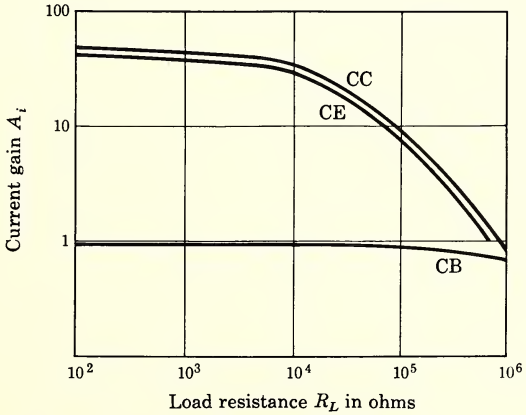


FIG. 7–7. Variation in current gain with load resistance.

1 megohm. For the CB stage, A_i has the maximum value of 0.98 (magnitudes only are shown on the curve), and this value is maintained as the resistance is increased, up to load resistances on the order of 10^5 ohms. However, a current gain of less than unity makes this stage unsuitable for use in a multistage current amplifier.

Figures 7–8 and 7–9 show that the input resistance of the CB stage is on the order of 100 ohms and that the output resistance is above 10^5 ohms. The CB stage may therefore be used as a buffer stage having a *low input resistance* and a *high output resistance.* The feedback term from output to input, h_{rb}, is the lowest value for the three circuits so good isolation between output and

input can be achieved. This stage does not introduce phase reversal, and, as will be shown later, it has a wider frequency response than the other two connections.

B. *Common collector circuit.* Figure 7–7 shows that it is possible to achieve a high current gain with one CC stage. However, a high current gain per stage cannot be achieved if CC stages are cascaded to form a multistage amplifier. The reason for this can be seen upon examination of Fig. 7–8. By comparing the dotted line representing $R_i = R_L$ and the curve of input resistance of the CC stage we see that the input resistance is higher than R_L by at least one order of magnitude up to $R_L = 10^5$ ohms. Thus for identical, cascaded CC

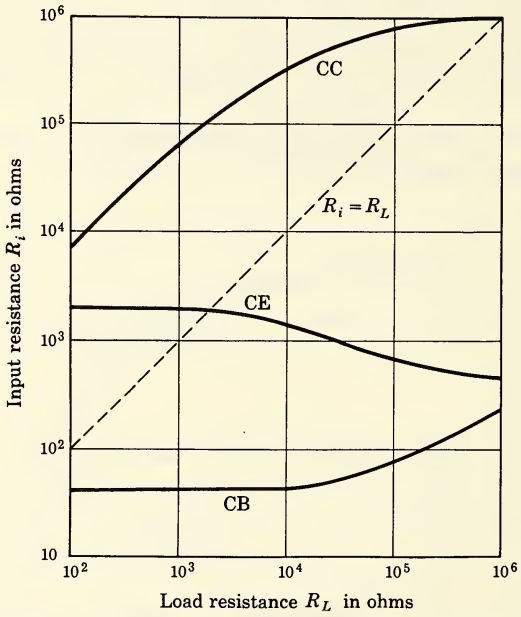


FIG. 7–8. Variation in input resistance with load resistance.

stages, the load resistance of a stage will bypass the input resistance of the next stage to such an extent that less than 10 per cent of the current will pass into the succeeding stage. This reduces the overall current gain of the stage to a value close to unity and makes the cascaded stages of little use.

One feature of the CC circuit is that the reverse feedback term, h_{ro}, is close to unity. This means that the output and input voltages will be approximately equal. Thus, however the input alternating voltage from base to collector alters, the output voltage from emitter to collector will follow. For this reason, the CC circuit is often known as the *emitter follower.* Although this circuit is unsuitable for cascading in a multistage amplifier, it does find a

use where its high input resistance and unity voltage gain characteristic can be utilized.

C. *Common emitter circuit.* The CE stage is seen to have a high current gain and input and output resistances which are intermediate between the values found for the CB and CC stage. When used in a cascaded amplifier, the output resistance of the CE stage given here is between 10^4 and 10^5 ohms, and the input resistance is about 2000 ohms. From Fig. 7–7, the current gain per stage in a cascaded amplifier where the input resistance of one stage forms the load of the previous stage approaches the short-circuit a-c value, h_{fe}. Unlike the CC stage, the load resistance can be made higher than the succeeding input resistance (2000 ohms), and so almost all of the alternating current leaving the stage enters the base lead of the following transistor.

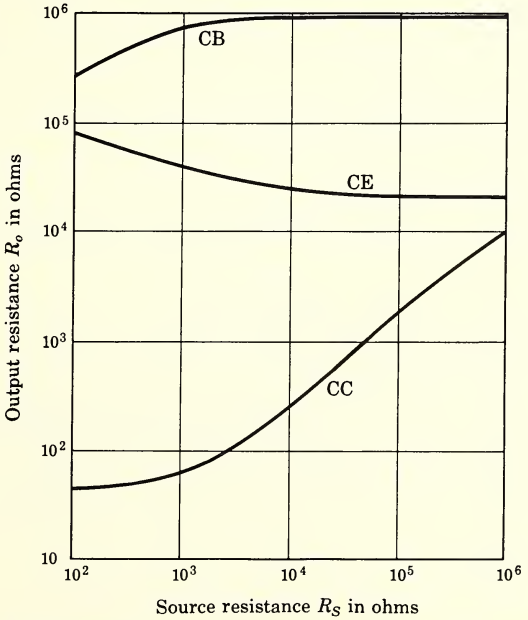


FIG. 7–9. Variation in output resistance with source resistance.

The CE stage introduces a phase reversal of 180° between output and input currents similar to a vacuum tube operated in the conventional manner. The frequency response of the CE stage is not as high as the CB stage, but, by selection of a suitable transistor, multistage amplifiers having a response up to hundreds of megacycles may be obtained.

7–5 Variation in *h* Parameters with I_E and V_{CE}

In general, the values of the *h* parameters are dependent on the quiescent operating point of the transistor. It is difficult to predict these variations with accuracy since there are many factors to take into account. Figures 7–10 and

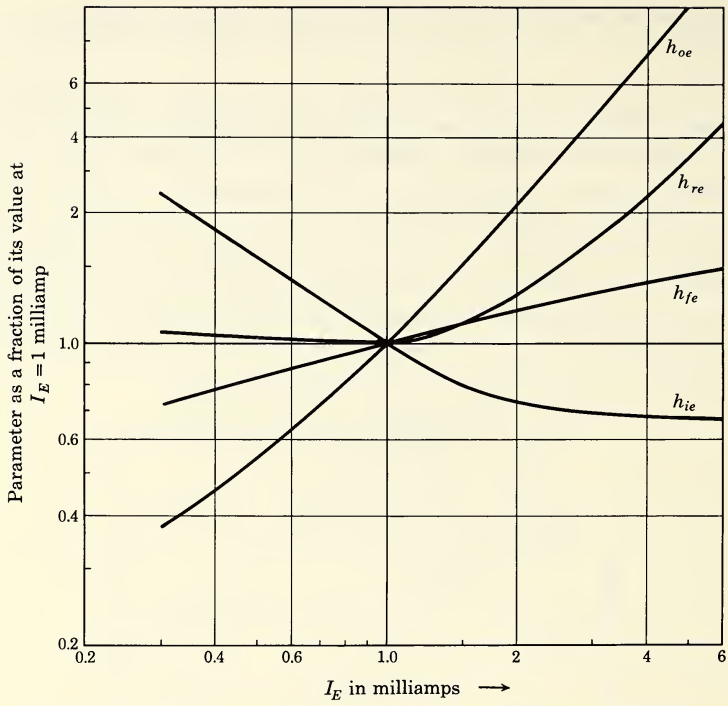


FIG. 7–10. An example of the variation in CE parameters with emitter current.

7–11 show h_{ie}, h_{oe}, h_{re}, and h_{fe} for an alloy-junction transistor as a function of I_E and V_{CE}. These curves were drawn from data supplied by the manufacturer. In Fig. 7–10, h_{fe} is shown to increase with I_E. For values of I_E higher than shown in this graph, h_{fe} would reach a maximum and decrease

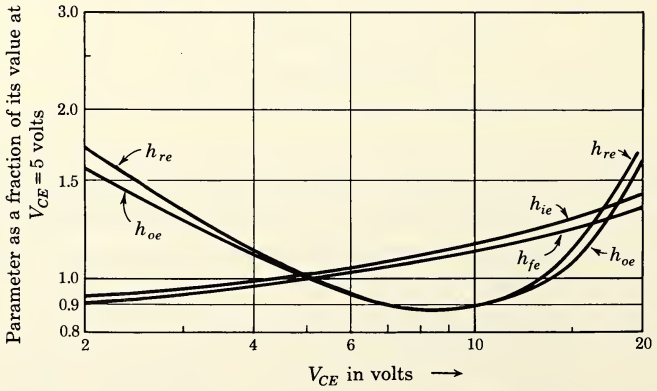


FIG. 7–11. An example of the variation in CE parameters with collector voltage.

(see Sec. 9–7). In Fig. 7–11, h_{ie} and h_{fe} steadily increase with V_{CE}, whereas h_{re} and h_{oe} have a minimum value at about $V_{CE} = 8$ volts. It is usually unwise to attempt to extrapolate curves of this nature much beyond the stated limits.

There are wide variations in the values of the parameters of transistors which have the same type number. There are also substantial variations in the form and the relative positions of the curves shown in Figs. 7–10 and 7–11. Manufacturers commonly specify only the average values of the h parameters at one operating point. If it is necessary to know the exact value of a particular transistor parameter, a measurement should be taken under the desired conditions.

BIBLIOGRAPHY

Fitchen, Franklin C., *Transistor Circuit Analysis and Design*, Princeton, N.J.: D. Van Nostrand Company, Inc., 1960

Joyce, Maurice V., and Kenneth K. Clarke, *Transistor Circuit Analysis*, Reading, Mass.: Addison-Wesley Publishing Company, Inc., 1961

Riddle, Robert L., and Marlin P. Ristenbatt, *Transistor Physics and Circuits*, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1958

PROBLEMS

7–1 A transistor has the following parameters: $h_{11} = 45$ ohms, $h_{12} = 4 \times 10^{-4}$, $h_{21} = -0.98$, $h_{22} = 10^{-5}$ mhos. From Table 7–3, find the parameters of the CE and CC circuits.

7–2 Using Fig. 7–2(c) as the equivalent circuit for the CC connection with a source resistance R_S and a load resistance R_L, write down the circuit equations and solve for the power gain.

7–3 Explain the advantages and disadvantages of using h parameters for transistor specification.

7–4 The transistor of Prob. 7–1 is operated in the CE connection with a load resistance R_L and with a resistance R_A between the collector and the base terminals. Find the current gain of the circuit.

7–5 Find the input resistance, the output resistance, and the current gain for a CE circuit with $R_S = 1000$ ohms and $R_L = 5000$ ohms. The CB specification of the transistor is $h_{ib} = 40$ ohms, $h_{rb} = 4 \times 10^{-4}$, $h_{fb} = -0.97$, $h_{ob} = 10^{-6}$ mhos.

7–6 A crystal microphone requires a load resistance of about one megohm and the output from the microphone is to be fed into a CE amplifying stage of input resistance 2000 ohms. Design a "buffer" stage to go

between the microphone and the amplifier. Use a transistor with specifications given in Table 7–4 and explain the reasons for your choice of circuit.

7–7 A transistor is operated as a CE amplifier with a resistance in the emitter lead. If the parameters of the transistor are h_{ie}, h_{re}, h_{fe}, and h_{oe}, find the h parameters of a two port network which includes the transistor and the resistance.

7–8 Two CE stages are coupled in series so that all of the alternating current from the collector of the first transistor flows into the base lead of the second transistor. If the two transistors are identical and have h parameters h_{ie}, h_{re}, h_{fe}, and h_{oe}, what are the h parameters of the two port network which includes both transistors?

7–9 Explain why CB or CC stages are never cascaded in a multistage amplifier. What are the chief features of a CE multistage amplifier?

7–10 Show that for the CE configuration:

$$h_{ie} = r_b + \frac{r_e}{1 - \alpha}, \qquad h_{re} \approx \frac{r_e}{r_c - r_m}$$

$$h_{fe} = \frac{\alpha}{1 - \alpha}, \qquad h_{oe} \approx \frac{1}{r_c - r_m}$$

7–11 A transistor having h parameters as given in Table 7–4 is used in a CC circuit. If the load resistance is 10,000 ohms and the source resistance is 1 megohm, calculate R_i, R_o, A_i, A_v, and A_p.

8

Single and Multistage a-c Amplifiers

The CE transistor circuit, although not perfect, has considerable merit as an amplifier. It has a relatively high current gain, it is small and light, it requires low voltage supplies only, and it has a high efficiency. There are two undesirable features, however, which must be taken into account. The more important of these is the variation which is exhibited by almost all transistor characteristics and parameters as the temperature of the device changes. A second adverse property of present-day transistors is the large variation which exists

between characteristics of different units of the same type number. These two factors will be shown to determine the circuits used for obtaining the d-c bias to the transistor.

The operation of the transistor as an a-c amplifier in the manner indicated in the two preceding chapters is conditional upon obtaining and maintaining a suitable quiescent operating point. Thus it is desirable to have the transistor operating with the same a-c gain, and with the same direct current and voltage in its output circuit, even though the input and output transistor characteristics depend upon temperature. In this chapter, it is shown that operating point stability can be achieved by careful design of the d-c biasing circuit. We will limit the analysis to the two most-used circuits at low frequencies: (a) a small-signal CE amplifier where current gain is of the greatest importance, and (b) a CE amplifier which is designed for maximum power output and where current gain is of secondary importance.

8–1 The Characteristic Curves

Figure 8–1 shows the direct currents which flow in the three transistor leads. The transistor is assumed to be in the CE connection. From Chap. 5 it has been shown that the transistor satisfies the equations:

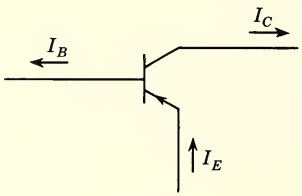


FIG. 8–1. Assigned directions of current.

$$I_C = I_E - I_B \tag{8–1}$$

$$I_C = \alpha I_E + I_{CBO} \tag{8–2}$$

and by eliminating I_E,

$$I_C = \frac{\alpha}{1 - \alpha} I_B + \frac{I_{CBO}}{1 - \alpha} \tag{8–3}$$

where I_{CBO} is the reverse-biased collector-junction current which flows when the emitter current is zero. We may write

$$\beta = \frac{\alpha}{1 - \alpha} \tag{8–4}$$

and

$$\frac{1}{1 - \alpha} = \beta + 1 \tag{8–5}$$

when Eq. (8–3) becomes

$$I_C = \beta I_B + I_{CEO} \tag{8–6}$$

where

$$I_{CEO} = (\beta + 1) I_{CBO}$$

If it is assumed that the quantity α is independent of I_C and that I_{CEO} can be neglected, Eq. (8–6) indicates that the collector current of a transistor is β

times the base current. In the small-signal two port network theory developed in Chaps. 6 and 7, it was shown that the small-signal short-circuit current gain of the transistor in the CE connection was h_{fe}. Hence, $\beta \approx h_{fe}$. However, h_{fe} is a function of I_C, V_{CB}, and temperature, and Eq. (8–6) relates to the d-c operation of the transistor so that β in this equation cannot simply be replaced by the small-signal value h_{fe}. Therefore, we will use the symbol h_{FE} (using capital subscripts) to denote the d-c or large signal value of β, where it is necessary to distinguish between the a-c and the d-c case. Similarly, h_{fb}

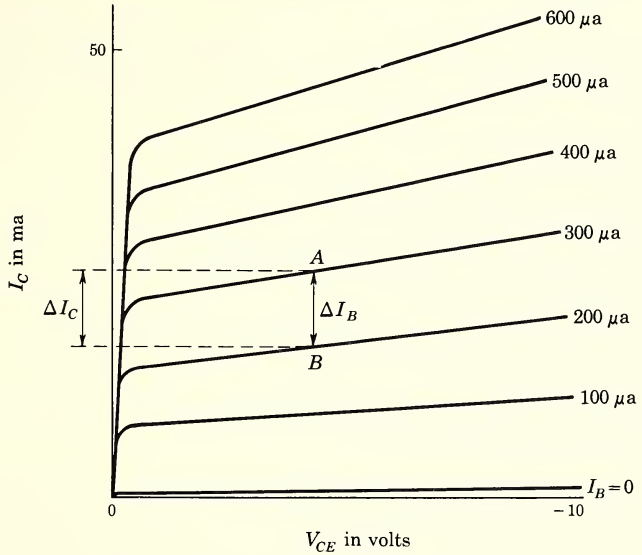


FIG. 8–2. Output characteristics of a typical *pnp* germanium alloy junction transistor.

which has been defined previously as the short-circuit small-signal current gain in the CB connection will be replaced by h_{FB} for d-c operation. The symbol α will be retained when no confusion will arise. Both α and β are terms in common usage in transistor circuitry.

The output characteristic curves of Fig. 8–2 can be explained by reference to Eq. (8–6). The curve labeled $I_B = 0$, represents the current I_{CEO}, a current $\beta + 1$ times I_{CBO}. If β and I_{CBO} were constant in Eq. (8–6) all the curves would be equally spaced, straight lines parallel to the V_{CE} axis. At low values of V_{CE}, such that the collector-to-base junction is just biased in the forward direction, the current falls to zero and transistor action ceases. In this region, Eq. (8–6) cannot be applied. Over the rest of the range, β and I_{CBO} are seen to be dependent on V_{CE} and I_C. The value of β (h_{FE}) can be calculated from the figure to be $\Delta I_C/\Delta I_B$ to give an average value between the points A and B.

There is little significant difference between the output characteristic of a silicon and a germanium transistor. For silicon, the value of I_{CBO} is much

smaller but the form of the curves is similar. The input characteristics of germanium and silicon transistors are different, however, as shown in Fig. 8–3. For both diagrams, the base current reverses for $V_{BE} \approx 0$. This is due to the flow of current in the base lead ($\sim -I_{CBO}$) when the collector to base is reverse biased. The input voltage required to produce $I_B > 0$ is considerably

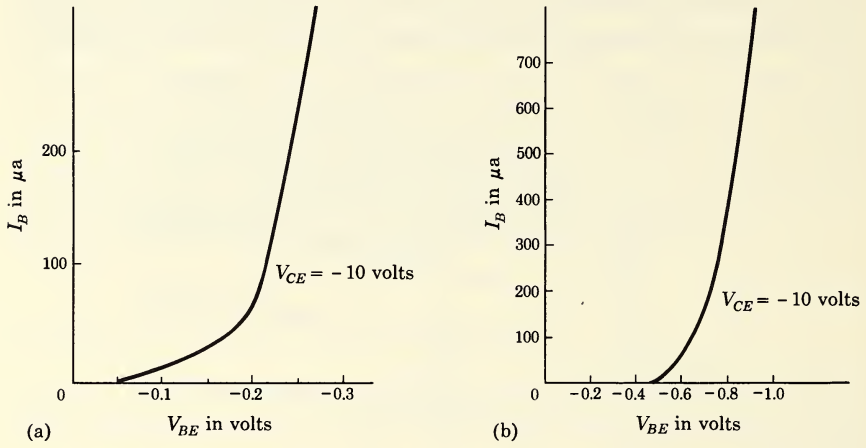


FIG. 8–3. Input characteristic of *pnp* alloy junction transistors: (a) germanium; (b) silicon.

different in the two types of transistors. It is of the order of 150 millivolts for germanium and 500 millivolts for silicon devices. The input curves lose their exponential form and become approximately straight lines at high values of base current. This occurs when the current is so high that most of the base-to-emitter voltage drop takes place across the ohmic resistance of the base.

8–2 The Load Line

Figure 8–4 shows the circuit of a CE amplifying stage. The supply battery V_{CC} is in series with the load resistance R_L and the collector-to-emitter terminals. This is similar to the output circuit of nearly all electronic devices. As far as alternating currents are concerned, the resistance of the supply can be ignored, and R_L is the load into which the transistor is working. I_C is dependent upon the supply voltage V_{CC}, the load R_L and the collector-to-emitter voltage V_{CE} as given in Eq. (8–7). If we designate all the voltages in

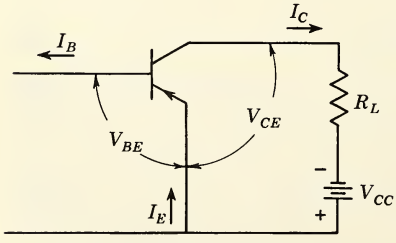


FIG. 8–4. The simple CE circuit.

the conventional manner with the terminal indicated by the first subscript as positive, we can write

$$V_{CC} = -I_C R_L + V_{CE} \tag{8–7}$$

where the supply voltage, V_{CC}, and the voltage, V_{CE}, are *negative* values for the *pnp* transistor. Equation (8–7) can be written in the form

$$I_C = \frac{V_{CE}}{R_L} - \frac{V_{CC}}{R_L} \tag{8–8}$$

and is plotted in Fig. 8–5 where, by convention for a *pnp* transistor, the *negative* values of V_{CE} are plotted. On these axes, Eq. (8–8) is the equation of a straight line of negative slope $(1/R_L)$, of intercept on the I_C axis of $-V_{CC}/R_L$,

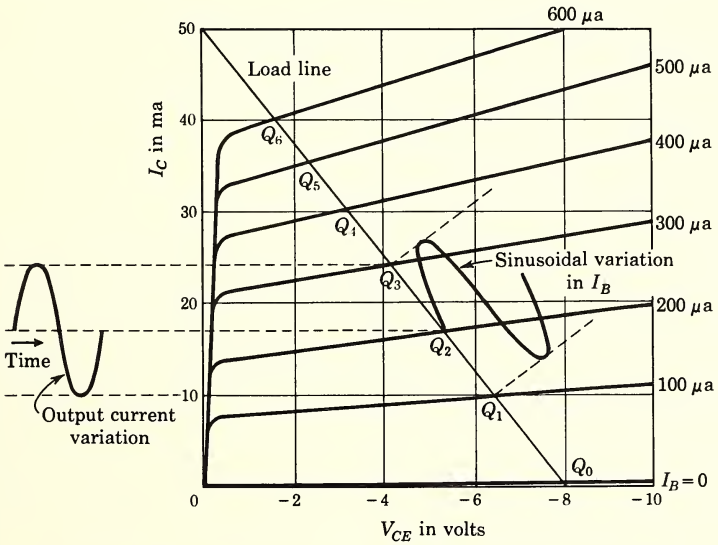


FIG. 8–5. The load line plotted on the output characteristic.

and which intercepts the $-V_{CE}$ axis at V_{CC}. The output characteristics shown are those of a medium power alloy-junction transistor with $V_{CC} = -8$ volts and $R_L = 160$ ohms. The straight line representing Eq. (8–8) is known as the *load line*, and its intersection with the output characteristics gives a set of quiescent operating points (Q points) for the device. Points Q_0, Q_1, Q_2, Q_3,... are obtained as I_B is changed. In the case shown in Fig. 8–5, the corresponding values of I_B are 0, 100, 200, 300 μa respectively. At all of the Q points, the circuit equation (8–8) is obeyed, and the device is operating on one of the characteristic curves shown in the diagram.

In Fig. 8–5, $I_B = 200$ μa has been fixed to give a Q point at Q_2. Then a sinusoidal variation of amplitude 100 μa has been impressed upon I_B. During this time, the circuit must still obey Eq. (8–8), and so the operating

point moves from Q_2 up to Q_3 down to Q_1 and back to Q_2 remaining on the load line all the time. The output current I_C as a function of time is shown to the left of the characteristic curves.

The selection of the Q point is decided upon by reference to the purpose of the stage. For a small-signal amplifier where power is to be conserved, a Q point would be chosen so that the quiescent value of I_C is as low as possible, provided that the output remains a faithful reproduction of the input signal. For an amplifier which is required to deliver a substantial amount of power into a load, the Q point should be selected so that the quiescent current is about one half of the maximum permissible collector current of the transistor. In this case, the allowable current swing above and below the Q point before distortion occurs is at a maximum (see Sec. 8–5).

The analysis which has been given in this section may be applied to the output circuit of many electronic devices. In the case of a transistor, however, the establishment of the quiescent operating point by fixing the base current results in a circuit having poor temperature stability. The reason for this is explained in the next section.

8–3 Operating Point Stability

It has already been shown in Part 1 of this book that many properties of semiconducting materials are dependent upon temperature. A transistor is usually operated so that its temperature is above that of its surroundings. The thermal capacity of the device is low because it is small and light. Thus the transistor will soon reach its operating temperature which is a function of its power dissipation, its thermal capacity, and the ambient conditions. Since many transistor parameters and characteristics are markedly temperature sensitive, the quiescent operating point should be established with the purpose of stabilizing the gain and output of the stage. The methods of achieving such a design will be explained. It will also be shown that, by poor design, the transistor can be made to heat itself up to such a temperature that it is destroyed. This process is known as *thermal runaway* and can be carried to completion only once for each transistor.

An increase in temperature has an important effect on three transistor quantities: I_{CBO}, α, and V_{BE}. In addition, there are likely to be significant differences in the measured values of I_{CBO} and α between transistors of the same type. The spread in the values of these quantities is troublesome in circuit design, and the manufacturers are continually trying to reduce it. It will be shown that circuits can be devised that will reduce the effect of changes in I_{CBO} or α whether these are the result of changes in temperature or of the replacement of the transistor by another of the same type.

The typical variation of I_{CBO} with temperature is given in Fig. 8–6 for silicon and germanium transistors. A considerable spread will be found in the values of I_{CBO} and in the slope of the curves when measurements are taken on

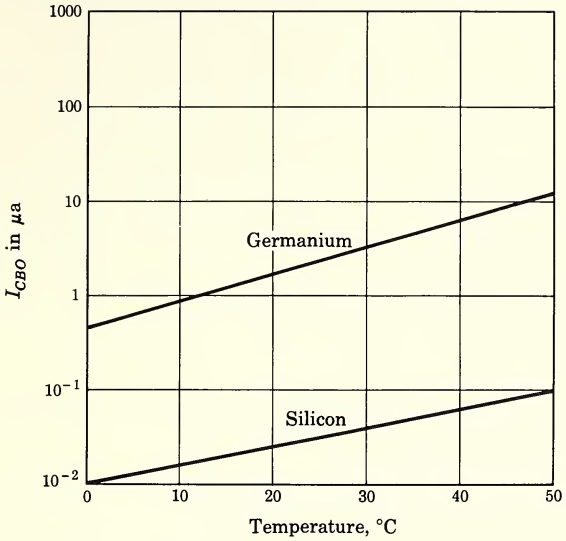


FIG. 8–6. Value of I_{CBO} for silicon and germanium transistors.

several types of transistors. This is because surface effects in the transistor and leakage currents across the junction contribute to I_{CBO} and make it dependent upon the manufacturing techniques. Transistors made of silicon have a much smaller value of I_{CBO} than those made of germanium because of the larger energy gap between the valence and conduction bands of the former material. In many cases, I_{CBO} for silicon transistors can be neglected without appreciable error being incurred even for high temperatures, and then the

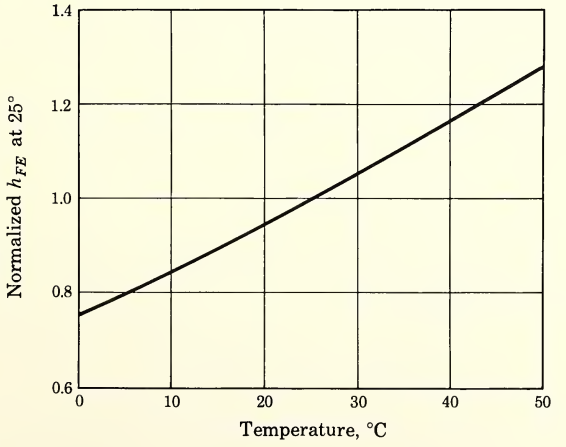


FIG. 8–7. Current gain (h_{FE}) as a function of temperature (normalized to 25° C).

variation in α or V_{BE} becomes of paramount importance. Measurements on both silicon and germanium transistors show that an increase in temperature of about 10° C results in I_{CBO} being doubled, whatever its initial value.

In a CE circuit, the quantity h_{FE} (the d-c value of β) determines the direct current gain. Furthermore, β [$= \alpha/(1 - \alpha)$] is more sensitive to changes in temperature than is α, and so the temperature variation in β is more likely to be given by the manufacturers. In Fig. 8–7, this value has been normalized to 25° C and is given only over the temperature range 0° to 50° C. Maximum and minimum values of h_{FE} (or possibly h_{fe}) for a given transistor at 25° C will usually be given in the data sheet supplied by the manufacturers. Thus a typical transistor may have the following specifications:

$$\left.\begin{array}{l} h_{FE} \text{ at } 25^\circ \text{ C} \\ I_C = 1 \text{ ma} \\ V_{CE} = -1 \text{ v} \end{array}\right\} 20 \text{ (minimum)} \quad 35 \text{ (typical)} \quad 60 \text{ (maximum)}$$

The base-to-emitter voltage necessary to produce a given emitter current for a germanium or a silicon transistor exhibits a temperature coefficient, and the voltage decreases at the rate of about 2.5 millivolts per degree centigrade. This is predictable from simple theory in the case of germanium, but the theory predicts a somewhat higher value for silicon which is not found in practice. There is little variation in this temperature coefficient between different transistors. For *pnp* transistor operation, the base-to-emitter voltage is negative. Thus, as the temperature rises, V_{BE} becomes less negative, and so the temperature coefficient is a positive quantity.

From Figs. 8–6 and 8–7 and from the information in the last paragraph, we may summarize the variation in I_{CBO}, h_{FE}, and V_{BE} over the temperature range 0° to 50° C in the following three equations.

$$\frac{\Delta I_{CBO}}{\Delta T} \sim 0.07 I_{CBO}\Big|_{T°C} \text{ amps } °\text{C}^{-1} \tag{8–9}$$

Equation (8–9) states that there is a change in I_{CBO} of the order of 7 per cent per °C of its value *at the temperature under consideration.*

$$\frac{\Delta h_{FE}}{\Delta T} \sim 0.012 h_{FE}\Big|_{25°C} °\text{C}^{-1} \tag{8–10}$$

Equation (8–10) states that there is a change in h_{FE} of the order of 1.2 per cent per °C *of its value at* 25° C.

$$\frac{\Delta V_{BE}}{\Delta T} \sim 0.0025 \text{ volts } °\text{C}^{-1} \tag{8–11}$$

Equation (8–11) states that, for a *pnp* transistor, V_{BE} required for a constant emitter current becomes more positive at the rate of 2.5 mv per °C.

Note the difference in the temperature coefficients of I_{CBO} and h_{FE}. I_{CBO}

shows an exponential dependence upon temperature, and thus its temperature coefficient increases as the temperature rises. The temperature coefficient of h_{FE} is constant, and thus h_{FE} at any temperature may be referred to its 25° C value. For the temperature range 25° to 35° C, I_{CBO} approximately doubles (i.e., shows an increase of 100 per cent), and h_{FE} increases by 12 per cent. It will be concluded, therefore, that as the temperature increases, I_{CBO} increases more and more rapidly with respect to the increase in h_{FE} as the Table 8–1 shows. These values are only approximate, but they apply to many small germanium and silicon alloy junction transistors.

TABLE 8–1. Variation in I_{CBO}, h_{FE}, and I_B normalized to 25°C and their typical values for a fixed-base biased transistor circuit.

	Normalized to 25° C			*Typical value for germanium at* 25° C	*Typical value for silicon at* 25° C
Temperature	25° C	35° C	45° C		
I_{CBO}	1	2	4	1 μa	0.01 μa
h_{FE}	1	1.12	1.24	50	50
I_B	1	1.025	1.05	100 μa	100 μa

In order to estimate the effects of temperature on the operation of the transistor, Eq. (8–12) will be used to represent the output characteristics in the CE connection. Here h_{FE} is used in place of β of Eq. (8–6).

$$I_C = h_{FE}I_B + (h_{FE} + 1)I_{CBO} \tag{8–12}$$

As a simplification, we shall assume an "idealized" transistor where I_{CBO}, h_{FE}, and I_B are independent of I_C and V_{CE}. The output characteristics of this transistor are straight lines parallel to the V_{CE} axis and are shown in Fig. 8–8(a) for room temperature. In this diagram, the value of $(h_{FE} + 1)I_{CBO}$ has been exaggerated for clarity. A load line has been drawn on the

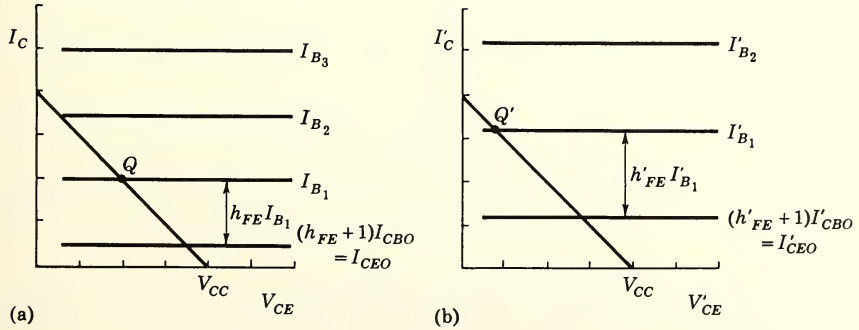


FIG. 8–8. Output characteristics of idealized transistor showing the load line (a) at room temperature; (b) at a higher temperature.

characteristics under the assumption that the transistor is connected in the circuit of Fig. 8–9. Now suppose that the temperature is increased so that the new currents and parameters are I'_C, I'_B, h'_{FE}, etc., related by the equation

$$I'_C = h'_{FE}I'_B + (h'_{FE} + 1)I'_{CBO} \tag{8–13}$$

The output characteristics at the new temperature are given in Fig. 8–8(b). Here $(h'_{FE} + 1)I'_{CBO}$ is considerably greater than its previous value, and also the spacing between the I'_{B1}, I'_{B2} curves has increased. As a result, the Q point shifts from Q to Q'.

The circuit of Fig. 8–9 is known as *fixed-base biasing*. If we designate all

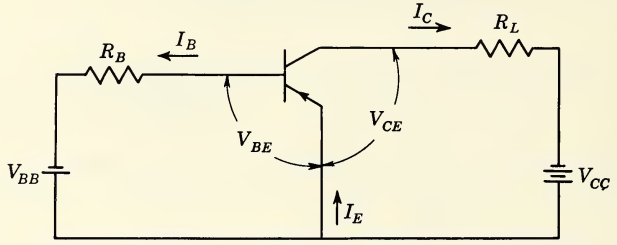


FIG. 8–9. The simplest d-c biasing circuit.

the voltages so that the terminal indicated by the first subscript is positive, we may write the following equations.

$$-I_B = \frac{V_{BB} - V_{BE}}{R_B} \tag{8–14}$$

and

$$-I'_B = \frac{V_{BB} - V'_{BE}}{R_B} \tag{8–15}$$

Thus,

$$\frac{\Delta I_B}{\Delta T} = \frac{I'_B - I_B}{\Delta T} = \frac{V'_{BE} - V_{BE}}{\Delta T\, R_B} = \frac{\Delta V_{BE}}{\Delta T\, R_B} = \frac{2.5}{R_B} \text{ ma } ^\circ\text{C}^{-1} \tag{8–16}$$

where R_B is in ohms, since $\Delta V_{BE}/\Delta T = 2.5$ mv per °C for a *pnp* transistor from Eq. (8–11). Thus I_B is also increased at the rate of $2.5/R_B$ ma per °C; an effect which still further increases I_C.

Under normal operating conditions, a rise in temperature of 20° to 40° C is common. I_{CBO}, therefore, increases to many times its initial value as the circuit warms up and increases in h_{FE} and I_B also occur. The shift in Q point shown in Fig. 8–8 usually makes this fixed-base biasing circuit unacceptable.

For the simple fixed-base biasing circuit of Fig. 8–9, variations in I_{CBO}, h_{FE}, and V_{BE} all contribute to an increase in I_C with temperature. Other circuits will be given which reduce this dependence of I_C upon temperature. Since I_{CBO} is the quantity which shows the greatest variation, the differential

coefficient $\partial I_C/\partial I_{CBO}$ is a useful measure of the temperature stability of a circuit. Defining this quantity as

$$\text{Stability factor, } S = \frac{\partial I_C}{\partial I_{CBO}} \tag{8–17}$$

we can avoid taking temperature measurements and yet have an indication of how the circuit will behave in practice. With this definition, *a high value of S indicates poor stability and a low value of S shows good stability.* The lowest value of S that can be obtained is unity since I_C must include I_{CBO}.

The stability factor of the circuit of Fig. 8–9 can be simply obtained since from Eq. (8–12),

$$S = \frac{\partial I_C}{\partial I_{CBO}} = h_{FE} + 1 \tag{8–18}$$

if I_B and h_{FE} are regarded as independent of I_{CBO}. This value of S is the highest that can be obtained showing that the circuit has very poor temperature stability.

The stability factor S gives a measure of the change in I_C with respect to I_{CBO} whether the change is the result of temperature variation or due to the transistor being replaced. For example, consider transistor A having $I_{CBO} = 1\ \mu a$ at 30° C and $h_{FE} = 49$ operating in a fixed-base bias circuit with the collector current $I_C = 0.5$ ma. In this circuit,

$$I_{CEO} = (h_{FE} + 1)I_{CBO} = 50\ \mu a$$

If the transistor were replaced by transistor B of the same type having the same value of h_{FE} but with $I_{CBO} = 5\ \mu a$ at the same temperature, I_{CEO} would rise to 250 μa and increase I_C by about 40 per cent. Changes as great as, or greater than, this are observed in practice for the fixed-base bias circuit where $S = h_{FE} + 1$. The performance of the stage in this respect is bad.

We have shown that I_{CBO} varies more widely with temperature than any other quantity, and so stabilizing with respect to I_{CBO} will adequately stabilize against temperature variations in h_{FE}. However, h_{FE} may be subject to variation irrespective of changes in I_{CBO}, such as when a transistor is replaced. We will use the symbol S_h to indicate this variation in I_C where,

$$S_h = \frac{\partial I_C}{\partial h_{FE}} \tag{8–19}$$

For the fixed-base bias circuit of Fig. 8–9,

$$I_C = h_{FE}I_B + (h_{FE} + 1)I_{CBO}$$

Therefore,

$$\begin{aligned} S_h = \frac{\partial I_C}{\partial h_{FE}} &= I_B + I_{CBO} \\ &= \frac{I_C - (h_{FE} + 1)I_{CBO} + h_{FE}I_{CBO}}{h_{FE}} \\ &= \frac{I_C - I_{CBO}}{h_{FE}} \end{aligned}$$

Considering small changes ΔI_C and Δh_{FE} in I_C and h_{FE} respectively, we may write

$$\frac{\Delta I_C}{\Delta h_{FE}} = \frac{I_C}{h_{FE}}\left(1 - \frac{I_{CBO}}{I_C}\right)$$

or

$$\frac{\Delta I_C}{I_C} = \frac{\Delta h_{FE}}{h_{FE}}\left(1 - \frac{I_{CBO}}{I_C}\right) \tag{8-20}$$

Equation (8–20) is a useful way of stating the variation in I_C as a result of changes in h_{FE}. It shows that for the fixed-base bias circuit, where $I_{CBO} \ll I_C$, the percentage change in I_C is the same as the percentage change in h_{FE}. Since the variation in h_{FE} between units may be +100 per cent to −50 per cent, this circuit is not acceptable.

8–4 Practical Biasing Circuits

The most commonly used CE biasing circuit is that of Fig. 8–10. In this circuit, a resistance R_E and a capacitance C_E are in parallel in the emitter lead. For frequencies much higher than given by

$$\frac{1}{2\pi f C_E} = R_E$$

the reactance of the capacitance is much smaller than the resistance R_E and may be neglected in practice. Thus for high frequencies, the stage reduces to the circuit of Fig. 8–9.

For direct currents, the resistance R_E forms a link between the input and output circuits since both I_B and I_C flow through this resistance to form the

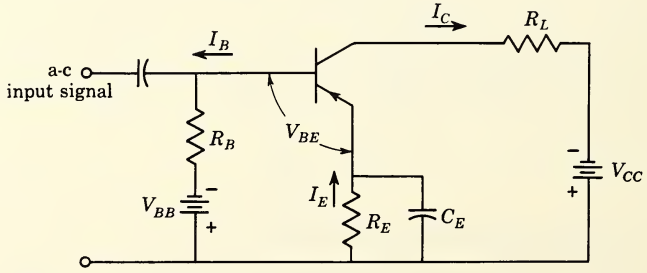


FIG. 8–10. CE biasing circuit with a resistance in the emitter lead

current I_E. Furthermore, the voltage developed across R_E by the output current opposes the input voltage from the battery supply V_{BB}. This is known as *current negative feedback* because the reverse voltage developed in the input circuit is proportional to the output current. Voltage negative feedback (see

Fig. 8–13) may be produced by feeding into the input circuit a proportion of the output voltage in such a way that it opposes the input voltage. The general properties of feedback circuits will not be discussed here but may be referred to in any general textbook on electronics. We will show, however, that the value of the stability factor can be reduced by this technique.

The following equations apply to the circuit of Fig. 8–10. For the transistor,

$$I_C = I_E - I_B \tag{8–1}$$

$$I_C = h_{FE}I_B + (h_{FE} + 1)I_{CBO} \tag{8–12}$$

By mesh analysis of the base-to-emitter circuit,

$$V_{BB} = -I_E R_E - I_B R_B + V_{BE} \tag{8–21}$$

Substituting for I_E and I_B from Eqs. (8–1) and (8–12), Eq. (8–21) becomes

$$V_{BE} - V_{BB} = I_C\left[R_E + \frac{R_B + R_E}{h_{FE}}\right] - \left[\frac{h_{FE} + 1}{h_{FE}}\right]I_{CBO}(R_B + R_E) \tag{8–22}$$

If the variations in V_{BE} and h_{FE} are ignored, Eq. (8–22) may be differentiated with respect to I_{CBO} to find the stability factor.

$$0 = S\left[R_E + \frac{R_B + R_E}{h_{FE}}\right] - \left[\frac{h_{FE} + 1}{h_{FE}}\right](R_B + R_E)$$

Therefore,

$$S = \frac{(h_{FE} + 1)(R_B + R_E)}{(h_{FE} + 1)R_E + R_B} = \frac{R_B + R_E}{R_E + \dfrac{R_B}{h_{FE} + 1}} = \frac{1 + R_B/R_E}{1 + \dfrac{R_B}{(h_{FE} + 1)R_E}} \tag{8–23}$$

When $R_E \to 0$, $S \to h_{FE} + 1$

When $R_B \to 0$, $S \to 1$

When R_E is zero, the circuit becomes identical with the previous circuit given in Fig. 8–9. When R_B is zero and R_E is finite, this circuit becomes a common base configuration as far as direct currents are concerned. Assuming V_{BE} is constant, I_E is fixed by V_{BB} and R_E and is not temperature dependent. Hence I_C can only increase as much as I_{CBO} increases. This results in $S = 1$. The circuit remains a CE amplifier for alternating currents since the signal is fed to the base terminal, and C_E bypasses R_E at signal frequency.

In practice, a compromise between the two extreme values of S is chosen. If R_E is made very high, it will dissipate too much power from the collector supply and will create an excessive voltage drop between the emitter and the common lead. If R_B is made too low it will act as a shunt to the input signal current and reduce the effective gain of the stage. For low power stages, where the rise in temperature is small, $S = 10$ can often be tolerated. For a typical value of $h_{FE} = 50$, this requires that $R_B/R_E \approx 11$ from Eq. (8–23).

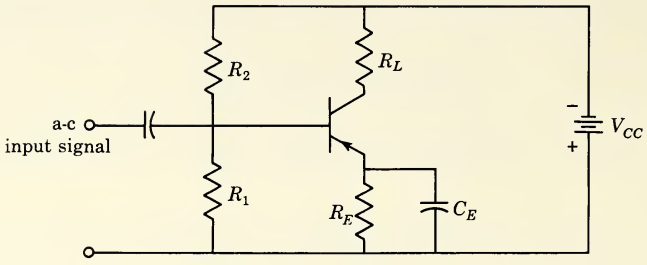


FIG. 8–11. A practical biasing circuit.

The variation in I_C because of changes in h_{FE} will now be investigated. Differentiating Eq. (8–22) with respect to h_{FE} gives

$$0 = S_h R_E + (R_E + R_B)\left(\frac{h_{FE}S_h - I_C}{h_{FE}^2}\right) - I_{CBO}\frac{(R_E + R_B)}{h_{FE}^2}[h_{FE} - (h_{FE} + 1)]$$

Thus,

$$S_h = \frac{\partial I_C}{\partial h_{FE}} = \frac{(I_C - I_{CBO})}{h_{FE}^2}\frac{(R_E + R_B)}{\dfrac{h_{FE}R_E + R_E + R_B}{h_{FE}}} = \frac{(I_C - I_{CBO})}{h_{FE}}\frac{S}{(h_{FE} + 1)}$$

Considering small changes in I_C and h_{FE},

$$\frac{\Delta I_C}{I_C} = \frac{\Delta h_{FE}}{h_{FE}}\frac{S}{(h_{FE} + 1)}\left(1 - \frac{I_{CBO}}{I_C}\right) \tag{8–24}$$

This represents an improvement by the factor $S/(h_{FE} + 1)$ over the performance of the fixed-base bias circuit of Fig. 8–9 (Eq. 8–20).

To avoid having two voltage supplies, the base current can be obtained from the collector supply V_{CC}. Figure 8–11 shows a practical circuit which is commonly used to achieve this. Resistances R_1 and R_2 form a voltage divider across the collector supply and can be replaced by the Thévenin equivalent consisting of a battery of voltage $V_{CC}R_1/(R_1 + R_2)$ and a series resistance of $R_1R_2/(R_1 + R_2)$ as shown in Fig. 8–12. With reference to the values used previously in this section,

$$V_{BB} = \frac{V_{CC}R_1}{R_1 + R_2}$$

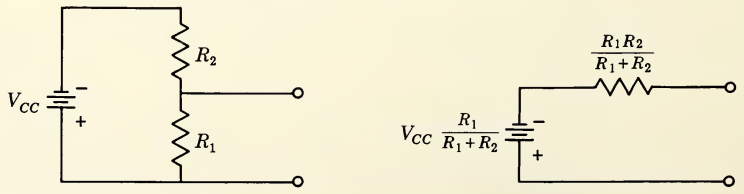


FIG. 8–12. Two circuits which are equivalent.

and $$R_B = \frac{R_1 R_2}{R_1 + R_2}$$

The circuit shown in Fig. 8–13 employs both current and voltage feedback. As before, a resistance is placed in the emitter lead to give current feedback under d-c conditions. The resistance R_F feeds back output voltage from the collector circuit to the base and is responsible for negative voltage feedback. Writing the d-c equation for the circuit, we have

$$V_{CC} = -I_E(R_E + R_L) - I_B R_F + V_{BE} \tag{8–25}$$

Substituting for I_E and I_B in terms of I_C and I_{CBO} from Eqs. (8–1) and (8–12),

$$I_C = \frac{(V_{BE} - V_{CC})h_{FE}}{R_F + (R_E + R_L)(h_{FE} + 1)} + \frac{I_{CBO}(h_{FE} + 1)(R_E + R_L + R_F)}{R_F + (R_E + R_L)(h_{FE} + 1)}$$

Thus,

$$S = \frac{\partial I_C}{\partial I_{CBO}} = \frac{(h_{FE} + 1)(R_E + R_L + R_F)}{R_F + (R_E + R_L)(h_{FE} + 1)} \tag{8–26}$$

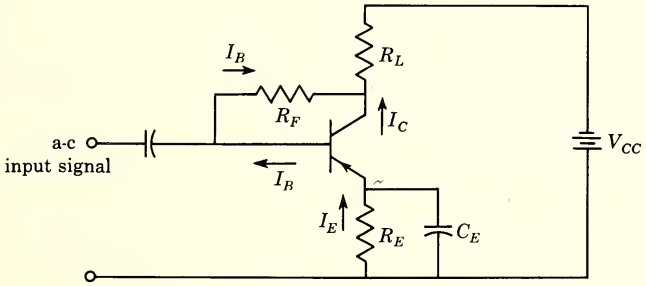


FIG. 8–13. A biasing circuit with current and voltage feedback.

If $R_F \gg (h_{FE} + 1)(R_E + R_L)$,

$$S \to h_{FE} + 1$$

If $R_F \ll (h_{FE} + 1)(R_F + R_L)$,

$$S \to 1 + \frac{R_F}{R_E + R_L}$$

When $R_E = 0$ in Fig. 8–13, only voltage feedback is being applied. It will be seen that R_F provides the negative voltage feedback, and it also sets the base current bias. Consider the case where the Q point is established so that $V_{CE} = V_{CC}/2$. Neglecting I_{CBO} and V_{BE}, the following equations apply.

$$I_B = \frac{-V_{CE}}{R_F} = \frac{-V_{CC}}{2R_F}$$

$$I_B(h_{FE} + 1) = I_E$$

and $$I_E = \frac{-V_{CC}}{2R_L}$$

Hence,

$$\frac{R_F}{R_L} = \frac{I_E}{I_B} = (h_{FE} + 1) \tag{8-27}$$

In a circuit containing a temperature sensitive device, it is usual to position the Q point at room temperature at about the midpoint of the load line as in this example since the maximum variation of the Q point in either direction is permitted. The stability factor is restricted to the value given by Eq. (8–26) with $R_E = 0$ and $R_F/R_L = h_{FE} + 1$. Thus,

$$S = \frac{h_{FE} + 2}{2} \approx \frac{h_{FE}}{2} \tag{8-28}$$

This value is high and so this simple circuit is rarely used.

Figure 8–14 shows one way in which the stability factor can be reduced while keeping the base current at its previous value. An additional supply,

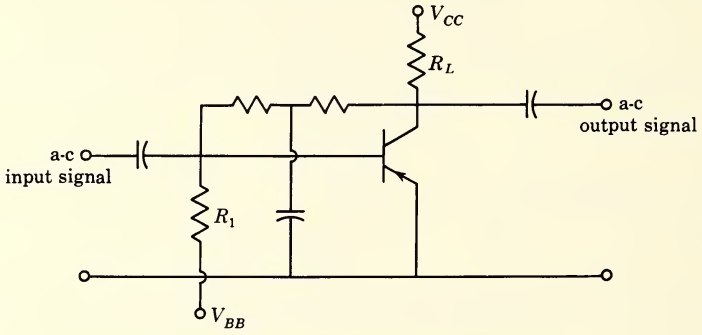


FIG. 8–14. A bias stabilization circuit using voltage feedback.

V_{BB} and a resistance R_1 are required as shown in the diagram. Notice also that R_F has been split and that the junction point is grounded for alternating currents. This prevents negative feedback from being effective at the signal frequency and so keeps the a-c gain of the stage independent of the applied feedback.

8–5 Power Amplifiers

The final stage of an amplifier is usually required to deliver power into a resistive load. The main purpose of the output stage is the conversion of power from the supply into useful a-c power in the load. The stage is designed to maximize power efficiency and to minimize transistor dissipation. Current and power gain are not major design factors since the overall gain requirements of the amplifier can be satisfied by the design of the preceding stages. For this reason, the design of an amplifier usually starts with a consideration of the output stage.

Figure 8–15 shows the circuit and the characteristic curves of a CE transistor amplifying stage. The load line corresponding to the series resistance R_L has been drawn on the characteristics. We distinguish three methods of operation of this type of stage according to the location of the quiescent point. Amplifying stages discussed so far have been operating in *class A*. In this condition, the transistor is operating during the whole of the input and output cycle and, if the device is linear, this results in distortion-free amplification. For power amplifiers, where maximum output is desired, the position of the Q point for class A operation will be approximately midway up the load line as shown in Fig. 8–15(a).

Class B operation is said to occur when the base circuit is biased so that very little base and collector current flows in the quiescent state. The transistor will then conduct only during one half of the input wave and will be

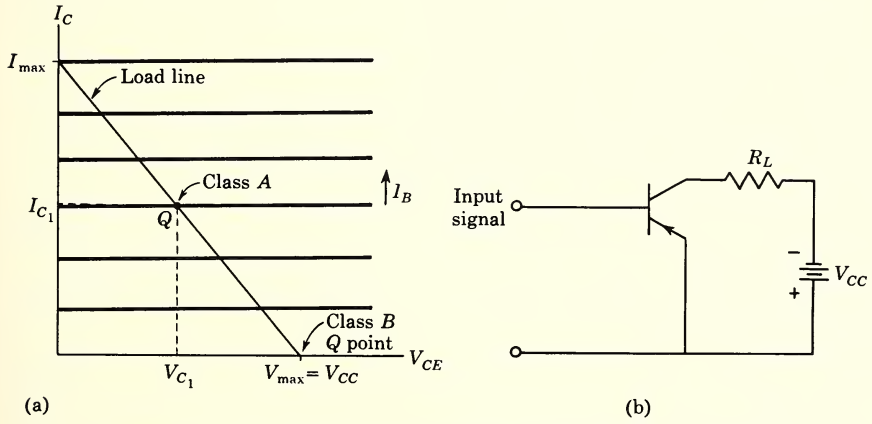


FIG. 8–15. (a) Idealized CE characteristics with load line. (b) CE transistor circuit with series resistive load (d-c bias not shown).

cut-off during the other half-cycle. Two transistors, operating out of phase with each other, are required to reproduce the original input signal. This is known as a "push-pull" circuit.

In *class C operation* of a single transistor stage, the base is biased so heavily that the transistor will conduct for less than 50 per cent of the signal cycle. The output consists of a reproduction of only a small fraction of the cycle of the input signal, and this introduces gross distortion. This type of operation is not suitable for untuned amplifiers but is used with high efficiencies for tuned amplifiers and oscillators.

If the Q point is positioned halfway up the load line as shown in the idealized case of a class A amplifier in Fig. 8–15(a), it will be located at

$$V_{C_1} = \frac{V_{max}}{2}, \qquad I_{C_1} = \frac{I_{max}}{2} \tag{8–29}$$

where V_{max} and I_{max} are the maximum permitted voltage and current for the transistor in the CE connection (see Sec. 8–6). When the maximum input signal is applied, the operating point swings up and down the load line so that the maximum current is I_{max}, and the minimum current is zero since $I_{CEO} = 0$ for the idealized case. Assuming a sinusoidal signal, the peak-to-peak current is I_{max} and

$$I_{rms} = \frac{I_{max}}{2\sqrt{2}} \tag{8–30}$$

The peak-to-peak swing of the a-c output voltage under these conditions is V_{max}, giving

$$V_{rms} = \frac{V_{max}}{2\sqrt{2}} \tag{8–31}$$

The a-c power delivered to the load is

$$P = I_{rms}V_{rms} = \frac{I_{max}V_{max}}{8} \tag{8–32}$$

Assuming that no distortion is produced, the power supplied to the stage is unchanged by the presence of the signal and is given by

$$P_S = I_{C_1}V_{CC} = \frac{I_{max}V_{max}}{2} \tag{8–33}$$

The efficiency of conversion of the d-c power from the supply to a-c power in the load is

$$\eta = \frac{P}{P_S} = 25\% \tag{8–34}$$

This is the maximum efficiency that can be obtained from a class A amplifier with a series resistive load.

TABLE 8–2. Distribution of power dissipation between transistor and load.

Circuit	*Condition*	*Transistor dissipation*	*Power in load*		*Total power supplied for a maximum transistor dissipation of 50 units*	*Efficiency*
			*a-c**	*d-c*		
Series resistive load	Quiescent	50	0	50	100	
	Maximum output	25	25	50	100	25%
Transformer-coupled load	Quiescent	50	0	0	50	
	Maximum output	25	25	0	50	50%

* Useful power output.

When the stage is quiescent, the voltage across the transistor is equal to the voltage across the load, and so half the power is dissipated in the transistor and half in the load. The total power supplied to the circuit and the d-c power dissipated in the load are unchanged when maximum signal is applied since the mean value of a sine wave is zero. Thus the transistor dissipation drops by the amount of the a-c power supplied to the load as shown in Table 8–2.

In the preceding analysis, the load has been chosen to give maximum power output. This requires a load of

$$R_L = \frac{V_{\max}}{I_{\max}} \tag{8–35}$$

If the maximum output is required, and yet the transistor has to work into a different load resistance, transformer coupling can be used. This circuit is shown in Fig. 8–16(b) where an external load R_L' is coupled by a transformer

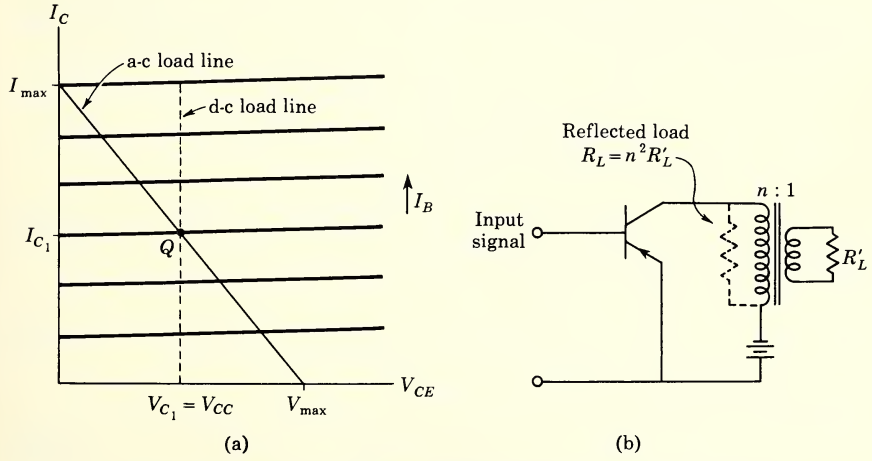


FIG. 8–16. (a) Idealized characteristics showing d-c and a-c load lines. (b) CE transistor circuit with transformer coupled load.

of turns ratio n to the power amplifier. The reflected load into the collector circuit is n^2R_L' which can be made equal to $V_{\max}/I_{\max}$ by the choice of a suitable value for n.

For direct currents, the resistance in the collector circuit is formed by the small d-c resistance of the primary windings of the transformer. The *d-c load line* is therefore almost vertical as shown in Fig. 8–16(a) and, for maximum output, the Q point is located on this line to give a quiescent current of $I_{\max}/2$. When the a-c signal causes the collector current to fluctuate, the operating point moves up and down the *a-c load line* shown on the diagram. The slope of this new line is $-1/R_L$, where R_L is the reflected resistance of the

load across the primary of the transformer. The a-c and d-c load lines must intersect at the Q point since this is the operating point for zero signal. Under a-c conditions, however, the apparent resistance in the collector circuit is that reflected across the transformer giving an a-c load line as shown.

The operating point will move along the a-c load line on both sides of the Q point. Thus the instantaneous voltage across the transistor will exceed the supply voltage, V_{CC}, during a portion of the cycle. The reason for this is that the change in current through the transformer produces a "back-emf" across the transformer primary winding which adds to the supply voltage during part of the cycle.

From Fig. 8–16(a) by considering similar triangles, the Q point is located at

$$I_{C_1} = \frac{I_{\max}}{2}, \qquad V_{C_1} = V_{CC} = \frac{V_{\max}}{2} \tag{8–36}$$

neglecting the small resistance of the transformer windings. When the transistor is operating at maximum capacity under class A conditions, the voltage and current swings are from $V_{\max}$ to zero and $I_{\max}$ to zero respectively. Hence,

$$V_{\text{rms}} = \frac{V_{\max}}{2\sqrt{2}} \quad \text{and} \quad I_{\text{rms}} = \frac{I_{\max}}{2\sqrt{2}} \tag{8–37}$$

and the useful a-c power in the load is

$$P = I_{\text{rms}} V_{\text{rms}} = \frac{I_{\max} V_{\max}}{8} \tag{8–38}$$

The total power supplied to the circuit is

$$P_S = V_{CC} I_{C_1} = \frac{V_{\max} I_{\max}}{4} \tag{8–39}$$

The power efficiency, η, for this circuit is

$$\eta = \frac{P}{P_S} = 50\% \tag{8–40}$$

Thus the maximum efficiency for the conversion of d-c power from the supply to a-c power in the load for the transformer-coupled class A amplifier is 50 per cent. The efficiency is double that of the previous case because no d-c power is dissipated in the load. Table 8–2 shows the power dissipation in the load and the transistor for this circuit assuming maximum transistor dissipation of 50 units.

Whether the load is in series with the collector or transformer coupled, Table 8–2 shows that the transistor dissipation drops to half its quiescent value when maximum signal is applied. This means that the transistor operating temperature is a function of the signal amplitude for both types of stage. Biasing circuits which stabilize the transistor stage against temperature variations, as discussed previously, are therefore a necessity.

8-6 Power Amplifiers: Practical Limitations

In an actual transistor, there are several limitations which must be taken into account. These are summarized here under the categories: maximum current, maximum voltage, maximum dissipation, minimum voltage (saturation).

The output characteristics of a typical transistor in the CE connection are shown in Fig. 8-17. For high values of collector current, the curves for

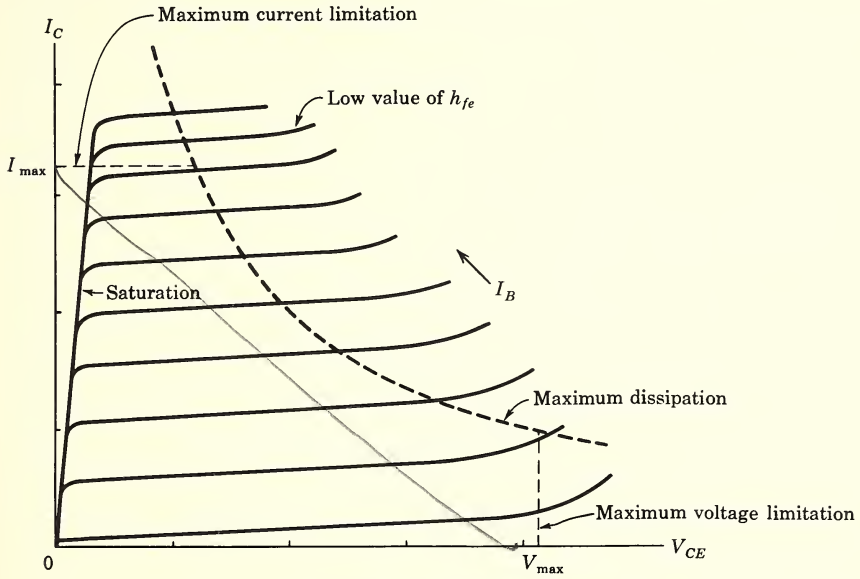


FIG. 8-17. Output characteristics showing limiting values.

constant base current increments are crowded closer together because of the drop in current gain as the collector current rises. Excessive distortion will be produced at peak currents if the collector current is raised indefinitely.

The maximum voltage limitation is determined by the onset of *collector multiplication*, *excessive surface leakage*, or *punch-through*. Collector multiplication is responsible for a marked rise in collector current as V_{CE} is increased. It is caused by an ionization avalanche phenomenon initiated by collisions of fast moving charge carriers in the collector-to-base depletion region. When a high collector-to-base voltage is present, a carrier gains enough energy in this region to generate additional electron-hole pairs which add to the collector current. Consideration of Fig. 8-17 shows that considerable distortion would be obtained if the transistor were operated in this region. Surface leakage is another undesirable effect in transistors and is also responsible for an unwanted increase in I_C as V_{CE} rises. Punch-through occurs when the depletion region at the collector end of the base is so large

that it meets up with the emitter-to-base depletion region and causes the active part of the base to disappear. The voltage at which this takes place is known as the *punch-through voltage* (see Sec. 9–11).

The transistor should not be operated with a power dissipation greater than a certain value specified by the manufacturers. For a given ambient temperature, the maximum dissipated power may be stated as $I_{C_1} V_{C_1} =$ constant. This equation has been plotted on Fig. 8–17, and the load line should always lie below this curve. A further discussion of this limitation is given in Sec. 8–7, Thermal Runaway.

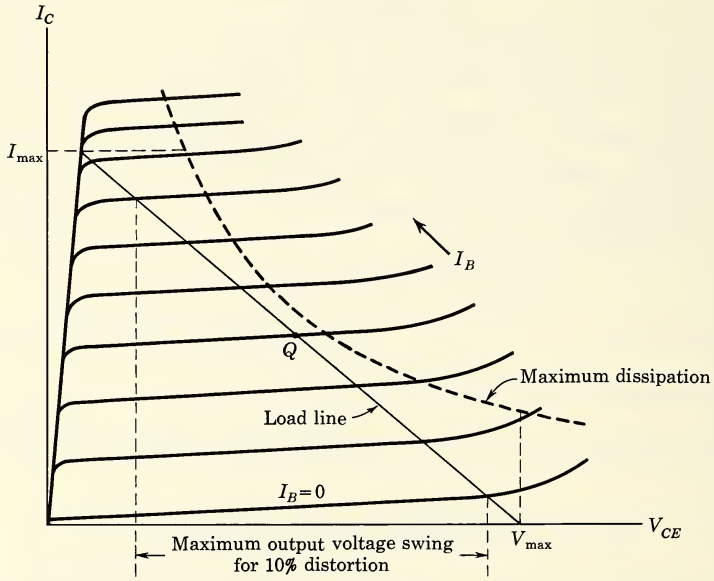


FIG. 8–18. Positioning of the load line showing Q point and maximum output voltage swing.

The part of the characteristics marked *saturation* forms another limitation for operation. Here, the collector-to-emitter voltage is insufficiently reverse biased for normal transistor operation. The voltage at which the transistor goes into saturation under given conditions determines the minimum allowable collector-to-emitter voltage for negligible distortion (see Sec. 10–7).

All these limitations must be taken into account by the circuit designer in deciding on a power amplifier circuit. For maximum power output and efficiency without too much distortion, the load and the collector supply voltage must be chosen to make good use of the output characteristics as shown in Fig. 8–18 for the class A single transistor circuit with a series resistive load.

A transistor operating as a small-signal class A amplifier introduces a

small amount of distortion into the amplified signal because of nonlinearities in the transistor characteristics. This distortion is usually less then five per cent and can be tolerated in many applications. When large-signal operation takes place, the distortion is increased, and it may prove to be the limiting

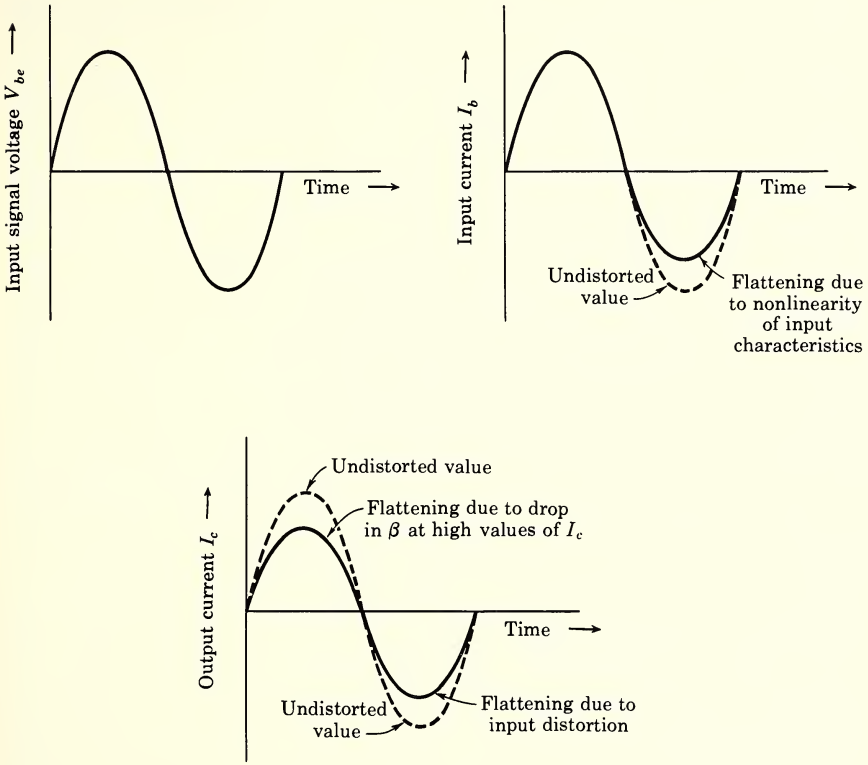


FIG. 8–19. Diagram showing effects of input and output distortion in a CE stage.

factor in the design of the output stage. It is found, in practice, that if a large-signal CE amplifier is driven from a low resistance source, distortion is minimized. This is the result of partial cancellation of the separate distortions occurring because of the nonlinearities in the input and output characteristic curves of the transistor. This is illustrated diagrammatically in Fig. 8–19.

8–7 Thermal Runaway

The junction temperature T_j of a transistor is determined by the ambient temperature T_A and the power dissipation P of the transistor according to the equation

$$T_j = T_A + KP \tag{8–41}$$

where K is known as the *thermal resistance*. From this equation it can be seen that

$$K = \frac{\Delta T_j}{\Delta P} \quad \text{or} \quad \frac{\Delta P}{\Delta T_j} = \frac{1}{K} \tag{8–42}$$

and so K is the junction temperature rise for unit increase in the power dissipation and has the dimensions °C (watts)$^{-1}$. Alternatively, $1/K$ can be expressed as the increase in power dissipation necessary to raise the junction temperature by 1° C. K is a function of the thermal conductivity of the materials used in the transistor, the shape and size of the transistor, and the method of mounting that is used.

If the collector current and voltage in a CE stage are fixed, the power dissipation is constant and is approximately $I_C V_{CE}$. The junction temperature T_j will attain its equilibrium value given by Eq. (8–41) within a few minutes of switching on the circuit, assuming that the ambient temperature remains constant. The maximum allowable power dissipation for the transistor at a given ambient temperature (usually 25° C) will usually be given by the manufacturers in the data sheet. If the ambient temperature is increased, the maximum dissipation must be reduced according to Eq. (8–41) and the curve of Fig. 8–20 applies. This curve shows the value of K.

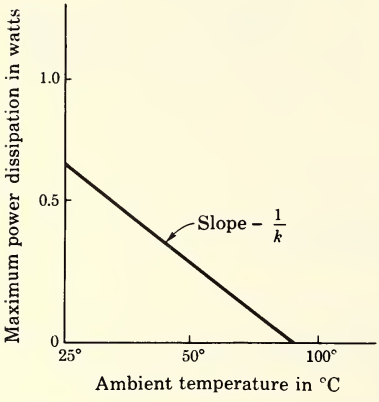


FIG. 8–20. Maximum power dissipation of a transistor as a function of ambient temperature.

Most CE amplifiers employ a biasing circuit which compensates for changes in quiescent operating point with temperature variation. These biasing circuits operate by negative feedback and inevitably introduce d-c power losses in the stage. In output stages where power efficiency is of importance, biasing circuits of only marginal stability are frequently used to reduce these losses, and considerable changes in I_C may occur as the temperature is increased. Under these conditions, the power dissipation, P, is no longer a constant but is a function of junction temperature, T_j. This function depends upon the properties of the transistor and also upon the external circuit.

It is possible to connect the transistor in a circuit so that the junction temperature rises continuously and eventually exceeds the safe limit. The effect is that a slight rise in temperature produces an increase in the power dissipation of the transistor, and this, in turn, increases the temperature of the junction. When the effect is cumulative, it is known as *thermal runaway* and is obviously undesirable.

From Eq. (8–42), the increase in the rate of transfer of energy from the junction (ΔP) for unit temperature rise ($\Delta T_j = 1°$ C) is $1/K$. This energy is transferred through the body, the case, and the mountings of the transistor and is a constant for a given type of mounting. However, the actual increase in energy dissipated in the transistor because of unit temperature rise is determined by the type of circuit and the characteristics of the transistor. If

$$\frac{\Delta P}{\Delta T_j} > \frac{1}{K} \tag{8–43}$$

more energy is dissipated in the transistor than can be transferred away from the junction, and so this is the condition for thermal runaway to occur.

For the typical circuit of Fig. 8–9,

$$\begin{aligned} P &= I_C V_{CE} \\ &= I_C(V_{CC} - I_C R_L) \end{aligned} \tag{8–44}$$

The condition to prevent thermal runaway, from Eq. (8–43) is

$$\frac{\Delta P}{\Delta T_j} \leqslant \frac{1}{K} \tag{8–45}$$

or, by differentiating Eq. (8–44),

$$\frac{\Delta I_C}{\Delta T_j}(V_{CC} - 2I_C R_L) \leqslant \frac{1}{K} \tag{8–46}$$

Equation (8–46) shows that for a given circuit, better stabilization of I_C, a larger load resistance, or a reduction in K might be used to prevent thermal runaway. $\Delta I_C/\Delta T_j$ may be expressed as $S(\partial I_{CBO}/\partial T_j)$, where $\partial I_{CBO}/\partial T_j$ is given in Fig. 8–6 and Eq. (8–9). Thus, reducing the stability factor, S, may be sufficient to prevent thermal runaway. The collector load R_L also plays a critical part in conditions for thermal runaway. Power output stages which are transformer coupled to the load are particularly susceptible to runaway because of the low d-c resistance of the transformer windings. Adding resistance in the collector lead, if it can be tolerated, may be sufficient to prevent this occurring. Finally, the value of K may be decreased by mounting the transistor with good thermal contact to a large, well-cooled metallic body known as a "heat-sink." The value of K when an infinite "heat-sink" is used is often given in the data sheet of the transistor.

8–8 Multistage Amplifier

Figure 8–21 shows the circuit of a CE two stage, small-signal amplifier where the two stages are identical. Transistors T_1 and T_2 are assumed to have the h parameters given in Table 7–4. Since the values of the parameters change with quiescent operating point, the emitter current and the collector voltage

will be assumed to have the values as specified for the transistor in the table. Typical variations in h parameters as the emitter current and collector voltage are varied are given in Figs. 7–10 and 7–11.

The performance quantities of the single stage have been given previously

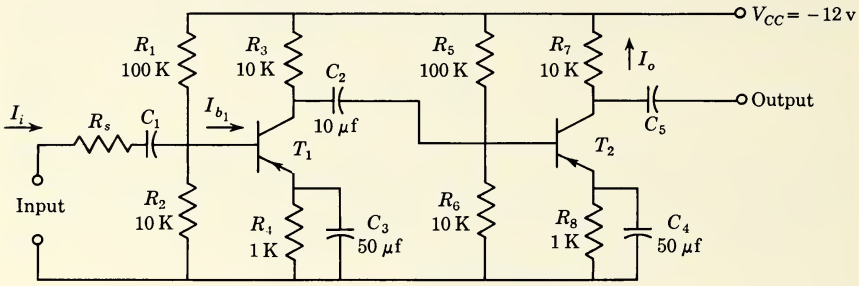


FIG. 8–21. A two-stage, small-signal amplifier.

in Table 7–2 in terms of a source resistance, R_S, and a collector load, R_L. From these tabulated values, the current gain of transistor T_2 is given by

$$A_{i_2} = \frac{h_{fe}}{1 + h_{oe}R_7} = \frac{49}{1 + (5 \times 10^{-5})10^4} \approx 32.5 \qquad (8\text{–}47)$$

In the multistage amplifier, however, the interstage coupling network must also be considered as shown in Fig. 8–22. I_1 and I_2 are the alternating currents supplied by the collector of T_1 and supplied to the base of T_2 respectively. The impedance of the V_{CC} supply to alternating currents is very low so R_5 and R_6 are effectively in parallel. Assuming that the reactance of C_2 can be ignored

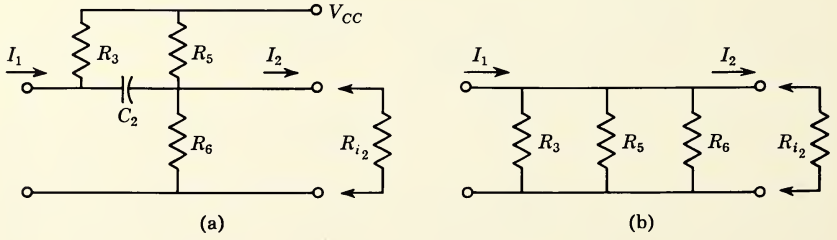


FIG. 8–22. (a) Interstage coupling network. (b) Equivalent circuit at high frequencies.

at the frequency of operation of the amplifier, the a-c equivalent circuit reduces to resistances R_3, R_5, and R_6 in parallel, as shown in Fig. 8–22(b). Writing R for the value of this parallel combination of resistances,

$$\frac{I_2}{I_1} = \frac{R}{R + R_{i_2}} \qquad (8\text{–}48)$$

For this circuit, $R = 4760$ ohms and R_{i_2} has still to be determined.

The load resistance, R_7, of the second stage is 10,000 ohms and so the input resistance of this stage is

$$R_{i_2} = \frac{\Delta^{he} R_7 + h_{ie}}{1 + h_{oe} R_7} = 1480 \text{ ohms} \tag{8–49}$$

from Table 7–4. Thus, from Eq. (8–48),

$$\frac{I_2}{I_1} \approx 0.77 \tag{8–50}$$

Transistor T_1 is operating into a load formed by the parallel combination of the four resistances R_3, R_5, R_6, and R_{i_2} as shown in Fig. 8–22. For the given values, the effective load is 1130 ohms. The current gain for transistor T_1 is, therefore,

$$A_{i_1} = \frac{h_{fe}}{1 + h_{oe} R_{L_1}} = \frac{49}{1 + (5 \times 10^{-5})1130} \approx 47 \tag{8–51}$$

In the input circuit, the current flowing into the base lead of the first transistor is only a fraction of the input current through R_S. The input resistance of T_1 is

$$R_{i_1} = \frac{\Delta^{he} R_{L_1} + h_{ie}}{1 + h_{oe} R_{L_1}} = 1920 \text{ ohms} \tag{8–52}$$

By analogy with Eq. (8–48), we can write

$$\frac{I_{b_1}}{I_i} = \frac{R'}{R' + R_{i_1}} \tag{8–53}$$

where R' is the parallel combination of R_1, R_2, and R_{i_1}. Now, $R' = 9100$ ohms, and so

$$\frac{I_{b_1}}{I_i} = \frac{9100}{9100 + 1920} \approx 0.82 \tag{8–54}$$

The total current gain through the circuit from the input lead to resistance R_7 is

$$\frac{I_o}{I_i} = 32.5(0.77)(47)0.82 = 960 \tag{8–55}$$

The stability factor for each transistor stage can be obtained from Eq. (8–23)

$$S = \frac{1 + R_B/R_E}{1 + \dfrac{R_B}{(h_{FE} + 1)R_E}} \tag{8–23}$$

h_{FE} will be specified by the manufacturers, and we will assume that it has the value 40. R_B is formed by R_1 and R_2 in parallel giving $R_B = 9100$ ohms. Hence,

$$S = \frac{1 + \dfrac{9100}{1000}}{1 + \dfrac{9100}{41(1000)}} = 8.3 \tag{8–56}$$

8–9 Frequency Response of the Multistage Amplifier

The frequency response of the typical multistage amplifier, shown in Fig. 8–21, is determined partly by the circuit elements and partly by the inherent frequency limitation of the transistors. At very low frequencies, the reactances of capacitances, C_1 and C_2 can no longer be ignored in the interstage coupling network, and they reduce the magnitude and affect the phase of the current which is flowing. The reactances of capacitances C_3 and C_4 also increase as the frequency is reduced and, as a result, the emitter resistances R_4 and R_8 are no longer adequately bypassed. This causes an additional decrease in gain.

At high frequencies, the current gain of the transistor decreases because of the time taken for the carriers to pass across the base region. The frequency at which the short-circuit CE current gain has dropped to 0.707 of its low frequency value is known as the *beta-cutoff frequency*, f_β. In the CB connection, the short-circuit current gain drops to 0.707 of its low frequency value at the *alpha-cutoff frequency*, f_α. The relationship between the two quantities is,

$$f_\alpha \approx h_{fe} f_\beta$$

and since h_{fe} is usually between 40 and 150, f_β is much less than f_α. A typical inexpensive alloy-junction transistor has an alpha-cutoff frequency of 5 Mc/sec and $h_{fe} = 50$. For this transistor, f_β is only 100 kc/sec and this constitutes a fundamental limitation on the use of this amplifier at high frequencies. (See Chap. 10 for a discussion of this point.)

In addition to this effect, the time constant formed by the collector load resistance and the output capacitance of the transistor produces a loss in gain at high frequencies. Because of this, the gain of the stage will have dropped to 0.707 of its low frequency value at a frequency given by

$$f = \frac{1}{C_{oe} R_L}$$

where C_{oe} is the collector output capacitance in the CE connection. More commonly, the output capacitance in the CB connection (C_{ob}) is specified. These two quantities are related according to

$$C_{oe} \approx h_{fe} C_{ob}$$

In many circuits, the beta-cutoff frequency is lower than the frequency at which the output time constant causes a loss in gain. When a high value of collector load, R_L, is used, however, the time constant may have to be considered.

The high frequency response of the CE stage is discussed in Secs. 10–2 through 10–5 where the hybrid-π model of the transistor is analyzed.

BIBLIOGRAPHY

Cote, Alfred J., Jr. and J. Barry Oakes, *Linear Vacuum-tube and Transistor Circuits*, New York: McGraw-Hill Book Company, Inc., 1961

Fitchen, Franklin C., *Transistor Circuit Analysis and Design*, Princeton, N.J.: D. Van Nostrand Company, Inc., 1960

Greiner, R. A., *Semiconductor Devices and Applications*, New York: McGraw-Hill Company, Inc., 1961

Hurley, Richard B., *Junction Transistor Electronics*, New York: John Wiley & Sons, Inc., 1958

Joyce, Maurice V. and Kenneth K. Clarke, *Transistor Circuit Analysis*, Reading, Mass.: Addison-Wesley Publishing Company, Inc., 1961

Pullen, Keats A., *Handbook of Transistor Circuit Design*, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1961

PROBLEMS

8–1 A transistor having the output characteristic shown in Fig. 8–5 is operated in the CE connection with a resistive load in the collector lead of 333 ohms and with $V_{CC} = -10$ volts. If $I_B = 200$ μa, what is the quiescent operating point? If an alternating current of 200 μa peak to peak is applied to the base, what is the value of the alternating output current? What is the maximum input alternating current that could be applied without excessive distortion?

8–2 Determine the value of h_{fe} for the quiescent operating point of the transistor in Prob. 8–1. If the input alternating current is 200 μa peak to peak, find the mean value of h_{oe} from a knowledge of A_i and h_{fe}.

8–3 Explain why the output characteristics of germanium and silicon transistors are about the same although there are significant differences in their respective input characteristics.

8–4 A *pnp* transistor is operated in the CE connection with a load resistance R_L in the collector lead. A high resistance R_1 is connected between the negative terminal of the collector supply battery and the base terminal. If $V_{CC} = -20$ volts and $\alpha = 0.98$, calculate the values of R_L and R_1 to make $V_{CE} = 10$ volts and $I_C = 2$ ma, making any reasonable assumptions considered necessary. What is the major limitation of this circuit?

8–5 For the circuit of Fig. 8–11 find (a) the stability factor, S, (b) S_h, (c) $\partial I_C/\partial V_{BE}$ given that $R_1 = 20{,}000$ ohms, $R_2 = 60{,}000$ ohms, $R_E = 2000$ ohms, $R_L = 3000$ ohms. Assume that $\alpha = 0.98$, $I_{CBO} = 3$ μa, $V_{BE} = -0.2$ volts, and $V_{CC} = -16$ volts.

ans. $S = 6.5$, $S_h = 4.3 \times 10^{-6}$ amps,
$\partial I_C/\partial V_{BE} = -4.25 \times 10^{-4}$ amps volts^{-1}

8–6 Using characteristic curves and equations where necessary, show how the location of the quiescent operating point is temperature sensitive.

8–7 A transistor is connected as in Fig. 8–11. If $V_{CC} = -20$ volts, $R_L = 2000$ ohms, $h_{FE} = 50$, and S is to be 12, calculate the values of the resistances to give as large an output voltage across R_L as possible without distortion. What is the minimum value for C_E if the circuit is to be used as an audio amplifier?

8–8 The transistor of Fig. 8–20 is operated in a CE connection where the stability factor S is 20 and there is no load resistance in the collector lead. $I_{CBO} = 10\ \mu$a at 25° C and $V_{CC} = -25$ volts. What is the critical value of collector current above which thermal runaway will occur?

ans. 13.2 ma

8–9 Two transistors are connected in cascade to supply a 1000 ohm load. Each has the following parameters: $h_{ib} = 30$ ohms, $h_{fe} = 49$, $h_{rb} = 10^{-4}$, $h_{ob} = 10^{-6}$ mhos. (a) Find the current and power gain of the circuit if both transistors are connected in the common base configuration. (b) Find the current and power gain if the first transistor is connected in a CB configuration and the second is in a CE connection. Design a suitable interstage coupling network in both cases.

PART 3

Further Theory of the Transistor, and the Tunnel Diode

9

Physical Characteristics of the Transistor

The continuity equation, which was derived in Chap. 4, is applied here to the base region of the transistor. First, a solution is found for the charge and current density as a function of distance in the base region when the transistor is operating under d-c conditions. This solution leads to a derivation of the transport factor and a discussion of the components of alpha. Then, the transistor is considered to have an applied alternating emitter current, and it is shown that the solution of the continuity equation under a-c conditions results in

an analogy with an RC transmission line. The transmission line is replaced by a lumped RC network, and this circuit leads to the useful *hybrid-π* representation of the transistor.

The effect of electric field in the base region is considered for high emitter currents in the diffusion transistor (Sec. 9–7) and for the important case of the drift transistor (Sec. 9–8). The chapter concludes with a brief review of other transistor characteristics.

9–1 Solution of the Continuity Equation in the Base Region for Direct Currents

The hole density in the base region of a *pnp* transistor is given by the continuity equation (4–9),

$$\frac{\partial p}{\partial t} = \frac{p_{n0} - p}{\tau_p} - \mu_p \frac{\partial}{\partial x}(p\mathscr{E}) + D_p \frac{\partial^2 p}{\partial x^2} \tag{9–1}$$

When only direct currents are flowing, $\partial p/\partial t$ is zero. For a transistor having abrupt junctions, where the resistivity of the base region is constant, the

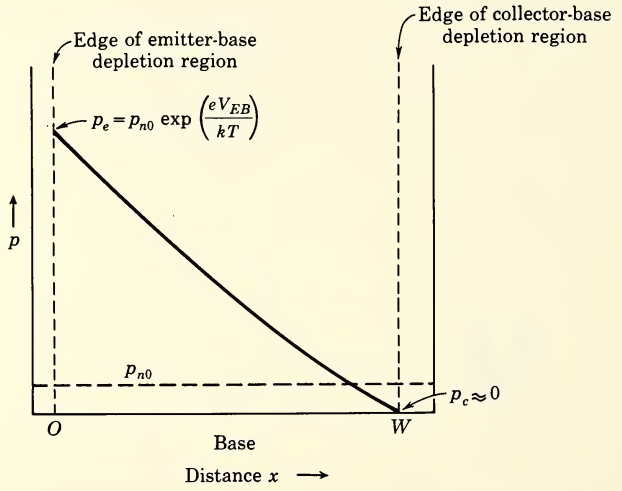


FIG. 9–1. Hole density in the base region of a *pnp* transistor.

passage of charge across the base under small-current conditions is the result of diffusion only, and $\mu_p \partial(p\mathscr{E})/\partial x$ is zero. The continuity equation then becomes

$$D_p \frac{d^2 p}{dx^2} = \frac{p - p_{n0}}{\tau_p} \tag{9–2}$$

or, since $D_p \tau_p = L_p^2$, where L_p is the diffusion length,

$$\frac{d^2 p}{dx^2} = \frac{p - p_{n0}}{L_p^2} \tag{9–3}$$

A solution of this equation is of the form,

$$p - p_{n0} = A \exp\left(-\frac{x}{L_p}\right) + B \exp\left(\frac{x}{L_p}\right) \tag{9-4}$$

where A and B are constants. These constants can be determined from the known values of p at the emitter-base and collector-base junctions. From Fig. 9–1, at the emitter-base junction,

$$x = 0, \qquad p = p_e = p_{n0} \exp\left(\frac{eV_{EB}}{kT}\right) \tag{9-5}$$

where $p_e \gg p_{n0}$ with $V_{EB} \gg kT/e$, and at the collector-base junction,

$$x = W, \qquad p = p_c \approx 0 \tag{9-6}$$

where $V_{CB} \ll -kT/e$.

At the collector end of the base region we note that both p_c and p_{n0} are very small quantities. From Eqs. (9–4) and (9–6), therefore,

$$A \exp\left(\frac{-W}{L_p}\right) + B \exp\left(\frac{W}{L_p}\right) = p_c - p_{n0} \approx 0$$

giving

$$A = -B \exp\left(\frac{2W}{L_p}\right) \tag{9-7}$$

Thus,

$$p - p_{n0} = -B\left[\exp\left(\frac{2W}{L_p}\right) \exp\left(\frac{-x}{L_p}\right) - \exp\left(\frac{x}{L_p}\right)\right] \tag{9-8}$$

At the emitter end of the base region, substituting Eq. (9–5) into Eq. (9–8),

$$p_e = -B[\exp\left(\frac{2W}{L_p}\right) - 1]$$

Hence,

$$B = \frac{p_e}{1 - \exp(2W/L_p)} \tag{9-9}$$

Substituting this value for B in Eq. (9–8)

$$p - p_{n0} = \frac{-p_e}{1 - \exp(2W/L_p)}\left[\exp\left(\frac{2W}{L_p}\right) \exp\left(\frac{-x}{L_p}\right) - \exp\left(\frac{x}{L_p}\right)\right] \tag{9-10}$$

The hole current density at any point in the base region is

$$J^+ = -eD_p \frac{dp}{dx}$$

Thus,

$$J^+ = \frac{-eD_p p_e}{L_p[1 - \exp(2W/L_p)]}\left[\exp\left(\frac{2W}{L_p}\right) \exp\left(-\frac{x}{L_p}\right) + \exp\left(\frac{x}{L_p}\right)\right] \tag{9-11}$$

This analysis is valid for direct currents and also for very low frequencies. We now define J_c^+ to be the hole current density leaving the base region at the base-collector junction where $x = W$. Similarly, J_e^+ is defined as the hole current density entering the base region at the emitter-base junction where $x = 0$. The ratio of these currents is known as the *transport factor*, β_0^*, where the subscript zero indicates that this is a low frequency value. From Eq. (9–11),

$$\beta_0^* = \frac{J_c^+}{J_e^+} = \frac{\exp(W/L_p) + \exp(W/L_p)}{\exp(2W/L_p) + 1}$$

$$= \frac{2}{\exp(W/L_p) + \exp(-W/L_p)} = \cosh(W/L_p) \quad (9–12)$$

It is desirable to make β_0^* as near to unity as possible since it is a principal component of alpha (see Sec. 9–2). This is achieved by making the base region so thin that $W \ll L_p$. In this case, $\exp(W/L_p)$, can be replaced by the first three terms of the series

$$1 + \frac{W}{L_p} + \frac{W^2}{2L_p^2} \cdots$$

and $\exp(-W/L_p)$ can be replaced by the first three terms of the series

$$1 - \frac{W}{L_p} + \frac{W^2}{2L_p^2} \cdots$$

Rewriting Eq. (9–12) in this approximate form,

$$\beta_0^* \approx \frac{2}{2 + W^2/L_p^2} = \frac{1}{1 + W^2/2L_p^2} \approx 1 - \frac{W^2}{2L_p^2} \quad (9–13)$$

9–2 Components of Alpha

The transport factor, β_0^*, which was derived in the previous section, is one of the three factors which determine the short-circuit current gain of the transistor at low frequencies, α_0. We can write,

$$\alpha_0 = \beta_0^* \gamma_0 \delta \quad (9–14)$$

where γ_0 is the *emitter efficiency* at low frequencies, and δ is the *collector multiplication factor*. From Fig. 9–2, the total current density entering the emitter is J_e, the hole current into the base region is designated J_e^+, and the electron current is J_e^-. The emitter efficiency as defined in Sec. 5–3B is $\gamma_0 = J_e^+/J_e$. Only the hole current, J_e^+, takes part in transistor action and this current is reduced by recombination in its passage across the base. The transport factor takes this recombination loss into account and so the hole current reaching the collector-to-base depletion region is $B_0^* J_e^+ = \beta_0^* \gamma_0 J_e$.

The *collector multiplication factor*, δ, is the ratio of the current leaving the collector region to the hole current entering from the base. This factor can

exceed unity, even for low voltage operation, in transistors where the resistivity of the collector region is high (for example, grown-junction types). The passage of a high current through the region results in an electric field of signal frequency, $\mathscr{E} = J_c^+ \rho$, where ρ is the resistivity of the material. This field modulates the thermally generated minority carrier current and produces a flow of electrons in the opposite direction to the hole flow. Thus, the collector current, J_c, is higher than the hole current from the base, J_c^+, and δ is greater than unity. δ will increase with temperature since the number of minority carriers rises with temperature.

When the collector-to-base voltage is high, an avalanche mechanism can develop which enhances the value of δ. Fast moving positive holes may be

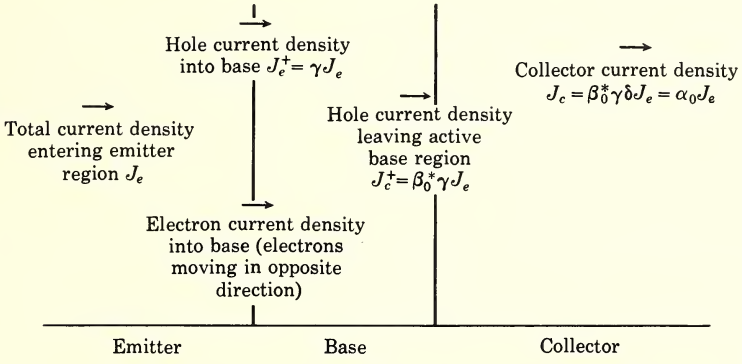


FIG. 9–2. Hole and electron currents flowing in the transistor as a result of an emitter current, J_e.

sufficiently energetic to produce secondary ionization in the collector depletion region and increase the total current. This increase in δ is a limiting factor in power transistor operation (see Sec. 8–6 and Fig. 8–17).

9–3 The Intrinsic Transistor

One of the purposes of transistor analysis is to devise a simple two port network, which will represent the device under given conditions. The transistor model can be split into two parts as shown in Fig. 9–3, (a) the *intrinsic transistor* and (b) the *external elements*. The intrinsic transistor, which is shown here as a "black-box," includes elements which represent the diffusion effects in the base region and other internal transistor characteristics. The external elements shown are the collector and emitter junction capacitances and a resistance, $r_{bb'}$. The two junction capacitances can be simply represented as two single capacitances connected to a common point on the active base region, B'. Between B' and the external base terminal B is the resistance $r_{bb'}$. This is known as the *base spreading resistance* and is the ohmic resistance of

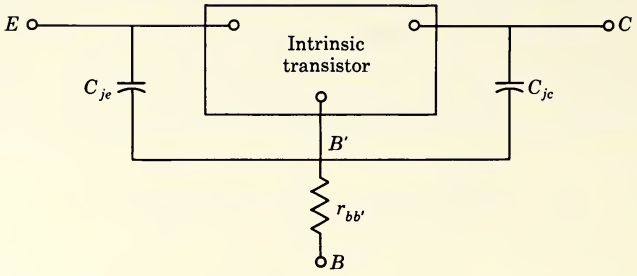


FIG. 9–3. Transistor model showing the intrinsic transistor.

the semiconducting material between the active base region and the base connection (see Fig. 9–4). There are also ohmic resistances associated with the emitter and collector leads, but these can usually be neglected. The base spreading resistance is of importance since it limits the high frequency and

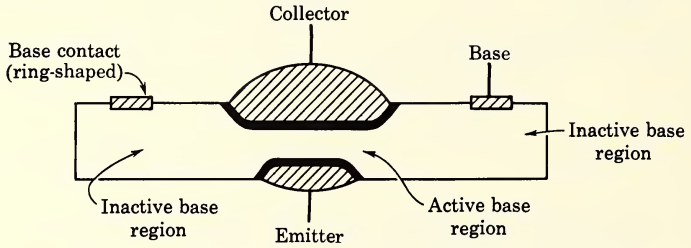


FIG. 9–4. Schematic diagram of an alloy-junction transistor showing the inactive part of the base region which is mainly responsible for the base spreading resistance.

pulse operation of the transistor (see Chap. 10), and the manufacturer will try to reduce it to as low a value as possible. Values for $r_{bb'}$ range from about ten to several hundred ohms.

9–4 The Continuity Equation in the Base Region for Small Alternating Signals

Consider that a small alternating voltage of value $V_{eb} \exp(j\omega t)$ is impressed on the direct emitter-to-base voltage V_{EB}. The instantaneous total voltage is then,

$$v_{eb} = V_{EB} + V_{eb} \exp(j\omega t) \tag{9–15}$$

Assuming $V_{EB} \gg kT/e$, the injected hole density in the base close to the base-to-emitter depletion region is given by

$$p_e = p_{n0} \exp\left\{\frac{e}{kT}\left[V_{EB} + V_{eb} \exp(j\omega t)\right]\right\}$$

$$= p_{n0}\left[\exp\left(\frac{eV_{EB}}{kT}\right)\right]\left\{\exp\left[\frac{e}{kT} V_{eb} \exp(j\omega t)\right]\right\} \qquad (9\text{–}16)$$

Putting in the condition that $V_{eb} \ll kT/e$, i.e., V_{eb} is restricted to well below 25.8 millivolts at 25° C, we can write $\exp[(e/kT)V_{eb} \exp(j\omega t)]$ in the approximate form $1 + (e/kT)V_{eb} \exp(j\omega t)$. Thus,

$$p_e = p_{n0} \exp\left(\frac{eV_{EB}}{kT}\right)\left[1 + \frac{e}{kT} V_{eb} \exp(j\omega t)\right]$$

$$= p_{n0} \exp\left(\frac{eV_{EB}}{kT}\right) + p_{n0}\left[\exp\left(\frac{eV_{EB}}{kT}\right)\right]\left[\frac{e}{kT} V_{eb} \exp(j\omega t)\right]$$

$$= p_{e_1} + p_{e_2} \exp(j\omega t) \qquad (9\text{–}17)$$

where p_{e_1} = direct component of hole density at $x = 0$ caused by voltage V_{EB},

p_{e_2} = maximum value of the alternating component of the hole density at $x = 0$ as a result of the alternating voltage V_{eb}.

When the transistor is operating with the collector-to-base junction reverse biased, the hole density at the collector end of the base region is approximately zero. Both the direct and alternating components of the hole density are reduced to zero at the collector junction and, in general, we can write the hole density at any point in the base region in the form

$$p = p_1 + p_2 \exp(j\omega t) \qquad (9\text{–}18)$$

where the subscript 1 indicates a direct quantity and the subscript 2 denotes an alternating quantity. Substituting Eq. (9–18) into the continuity equation (9–1), and neglecting the drift term,

$$j\omega p_2 \exp(j\omega t) = \frac{p_{n0} - p_1 - p_2 \exp(j\omega t)}{\tau_p} + D_p \frac{\partial^2 p_1}{\partial x^2} + D_p \frac{\partial^2}{\partial x^2}[p_2 \exp(j\omega t)] \qquad (9\text{–}19)$$

This equation can be split into two parts: a d-c equation which is similar to Eq. (9–2),

$$0 = \frac{p_{n0} - p_1}{\tau_p} + D_p \frac{\partial^2 p_1}{\partial x^2} \qquad (9\text{–}20)$$

and an a-c equation,

$$j\omega p_2 \exp(j\omega t) = -\frac{p_2}{\tau_p} \exp(j\omega t) + D_p \frac{\partial^2}{\partial x^2}[p_2 \exp(j\omega t)]$$

or, by dividing through by $\exp(j\omega t)$,

$$j\omega p_2 = -\frac{p_2}{\tau_p} + D_p \frac{\partial^2 p_2}{\partial x^2} \qquad (9\text{–}21)$$

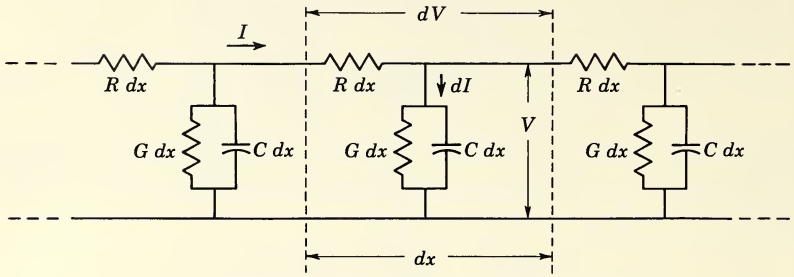


FIG. 9–5. Leaky RC transmission line.

Equation 9–21 is similar to the equation which can be derived for the RC transmission line shown in Fig. 9–5. The series resistive element has a resistance R per unit length of the line. The parallel resistance is most conveniently represented by its conductance G per unit length, and the capacitance has a value C per unit length. For an element of length dx, and for alternating currents and voltages defined in Fig. 9–5,

$$I = \frac{1}{R}\frac{dV}{dx} \tag{9–22}$$

and, by differentiating with respect to x,

$$\frac{dI}{dx} = \frac{1}{R}\frac{d^2V}{dx^2} \tag{9–23}$$

Now,
$$dI = (G + j\omega C)V\,dx$$

so
$$\frac{dI}{dx} = (G + j\omega C)V \tag{9–24}$$

Combining Eqs. (9–23) and (9–24),

$$j\omega CV = -GV + \frac{1}{R}\frac{d^2V}{dx^2} \tag{9–25}$$

The alternating charge on the capacitance is given by

$$Q = CV \tag{9–26}$$

Thus,
$$j\omega Q = -\frac{G}{C}Q + \frac{1}{CR}\frac{d^2Q}{dx^2} \tag{9–27}$$

Equation (9–27) is an equation for the maximum value of the alternating charge on the line as a function of distance. Equation (9–28) is an equation formed by multiplying all the three terms of the a-c continuity equation, Eq. (9–21), by the electronic charge, e.

$$j\omega ep_2 = -\frac{ep_2}{\tau_p} + D_p\frac{\partial^2(ep_2)}{\partial x^2} \tag{9–28}$$

Comparing Eqs. (9–27) and (9–28), it is seen that an analogy exists between the alternating charge in the base region of the transistor (ep_2) and the alternating charge on the line (Q).

The graphical solution of the two equations is shown in Fig. 9–6. The solution will not be derived here, but the form of it can be seen by considering the limiting values of the hole density at the two ends of the base region. At

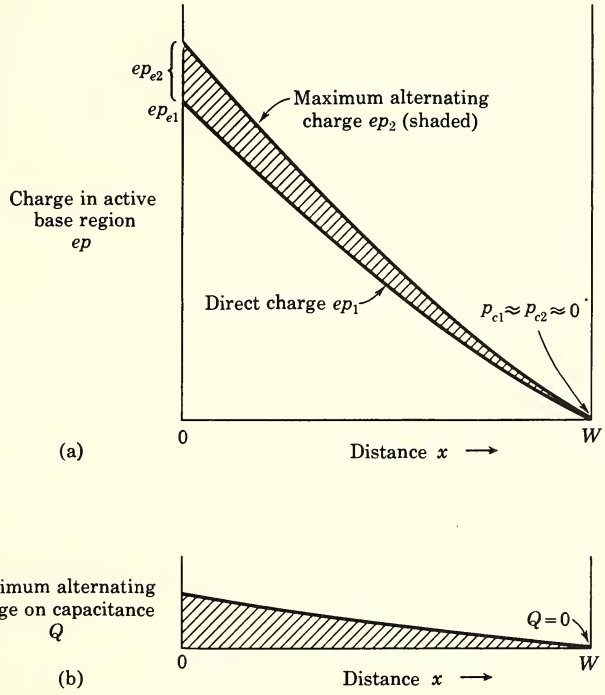


FIG. 9–6. (a) Charge in the active base region of the transistor showing maximum value of the alternating charge as a function of distance. (b) Maximum value of alternating charge on analogous transmission line as a function of distance.

the emitter-base junction, $x = 0$ and p_{e_1} and p_{e_2} are the direct and alternating components of hole density having the values given in Eq. (9–17). At the collector-base junction, $x = W$ and the direct and alternating values of the hole density, p_{c_1} and p_{c_2}, are both approximately zero. This assumes that the collector-to-base junction is reverse biased with $V_{CB} \ll -kT/e$. The analogous condition to this on the transmission line representation is that $Q = 0$ at $x = W$. The line is, therefore, short-circuited at its far end.

Between $x = 0$ and $x = W$ on the transmission line, the conductance, G, allows a fraction of the charge to leak away. Thus, the current flow through

G represents the loss of holes in the base by recombination. If recombination in the base is neglected, G must be made zero.

9–5 Base Width Modulation: Transmission Line Analogy

The previous section has shown that there is an analogy between the alternating charge at any point in the base region of a transistor and the alternating

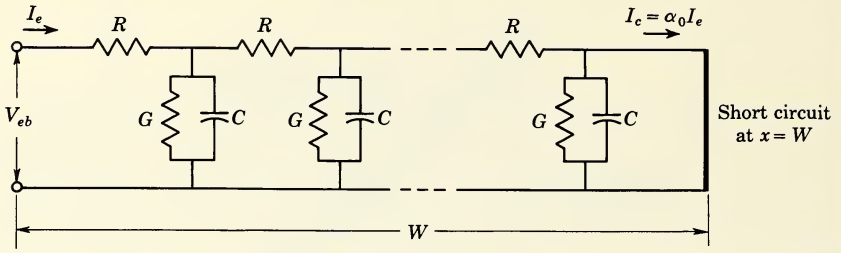


FIG. 9–7. Alternating currents and voltages on the line.

charge on the transmission line of Fig. 9–5. It has been shown that there is a short circuit at the position $x = W$ corresponding to the condition that the alternating hole density at the collector-to-base junction is zero. The analogy is carried further in Fig. 9–7 where the alternating currents and voltages which are known to be present in the transistor are applied to the line. The current entering the transmission line is the alternating emitter current, I_e, and the current passing through the short circuit at low frequencies is the

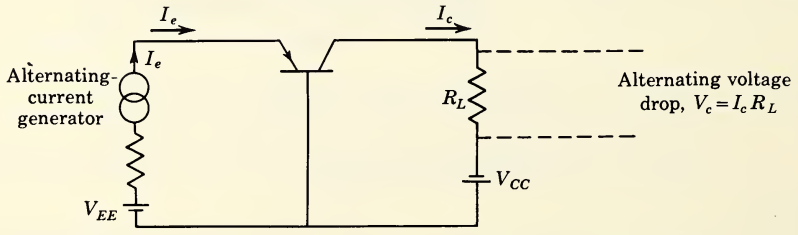


FIG. 9–8. Alternating voltage drop across load resistance, R_L.

collector current I_c. The amount of recombination in the base region is accounted for by the presence of the conductance, G, and so the collector current, I_c, can be written as $\beta_0^* I_e$. When the emitter efficiency and the collector multiplication factor are taken to be unity, then $\beta_0^* = \alpha_0$ and $I_c = \alpha_0 I_e$.

So far, it has been assumed that the collector-to-base voltage is a constant. This is not a valid assumption, in practice, since the presence of a load in the

collector circuit introduces a voltage drop between the collector supply and the collector terminal as shown for the CB circuit in Fig. 9–8. An alternating voltage between the collector and base terminals is produced here by the alternating collector current flowing through R_L. In general, therefore, we have to consider the effect of a collector modulating voltage on the transmission line analogy. This can be done by recalling that the collector-to-base voltage determines the depletion width at the collector end of the base region, and hence the active width, W, of the base region itself. In the analysis which follows, we shall distinguish between the active base, B', and the base terminal, B. The alternating voltage that is responsible for the change in base width is that between the collector and the active base and is designated $V_{cb'}$. The voltage, $V_{b'b}$, is across the base spreading resistance, $r_{bb'}$, of Fig. 9–3 and the

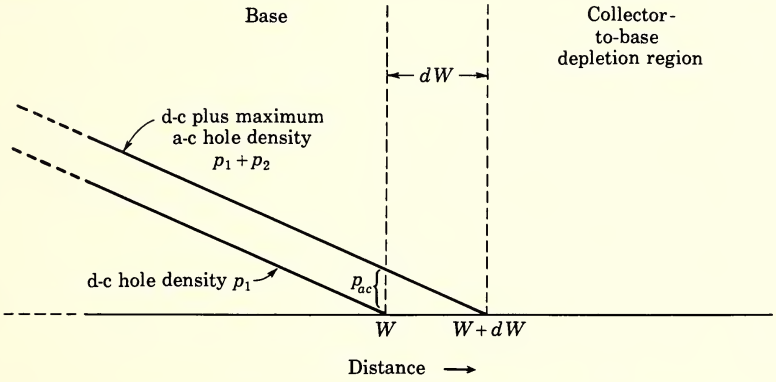


FIG. 9–9. Enlarged view of collector-to-base junction.

sum of these two voltages is V_{cb} which is across the external collector and base terminals. In some cases, $V_{b'b}$ can be neglected and $V_{cb'}$ and V_{cb} assumed to be identical. The correct notation will be retained in this section since the intrinsic transistor is being discussed.

In Fig. 9–9 a much enlarged view of the edge of the base-to-collector depletion region is shown. When no alternating emitter current is applied, the active region of the base ends at $x = W$ as shown. When the alternating current reaches its maximum value, the upper curve of hole density applies and the active region of the base extends out to $W + dW$. Making the assumptions, $dW \ll W$, $p_{e_2} \ll p_{e_1}$ and neglecting recombination, the slopes of the two hole density curves given in Fig. 9–9 are equal and are given by

$$\frac{dp}{dx} = -\frac{p_{e_1}}{W} \qquad \text{for the d-c case} \tag{9–29}$$

The boundary conditions,

$$p_{c_1} + p_{c_2} = 0 \qquad \text{at } x = W + dW \tag{9–30}$$

can be transformed to

$$p_2 = p_{ac} \qquad \text{at } x = W \tag{9-31}$$

by making use of the relationship

$$\frac{p_{ac}}{dW} = -\frac{dp}{dx} = \frac{p_{e_1}}{W}$$

Thus,

$$p_{ac} = \frac{p_{e_1}}{W}\, dW \tag{9-32}$$

The boundary condition of Eq. (9–30) means that the effective base width is a function of the collector voltage. The transformed boundary condition given in Eq. (9–31) shows that the active base width may be considered to be

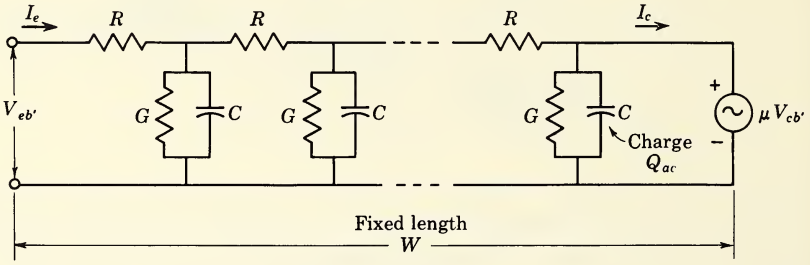


FIG. 9–10. Voltage generator, $\mu V_{cb'}$, representing the effects of base-width modulation on the transmission line.

fixed at the value W and then the hole density at $x = W$ is given by Eq. (9–32). Considering a junction of unit area, we can write the direct current entering from the emitter as

$$I_E = -eD_p \frac{dp}{dx}$$

$$= eD_p \frac{p_{e_1}}{W} \tag{9-33}$$

Hence,

$$p_{ac} = \frac{I_E}{eD_p}\, dW$$

$$= \frac{I_E}{eD_p} \frac{dW}{dV_{CB'}}\, dV_{CB'} \tag{9-34}$$

or

$$ep_{dc} = \frac{I_E}{D_p} \frac{dW}{dV_{CB'}}\, V_{cb'} \tag{9-35}$$

Returning to the transmission line analog, we see that the boundary condition of Eq. (9–30) leads to a transmission line with a length that is a

function of the collector-to-base voltage. This is an inconvenience which can be overcome by using the transformed boundary condition of Eq. (9–31) which fixes the line length at W and introduces a charge $Q_{ac} = ep_{ac}$ at $x = W$. This is equivalent to replacing the short circuit by a voltage generator of value $\mu V_{cb'}$, as shown in Fig. 9–10, where

$$\mu V_{cb'} C = Q_{ac} = ep_{ac} \tag{9–36}$$

or

$$\mu = \frac{ep_{ac}}{CV_{cb'}} = \frac{I_E}{CD_p} \frac{dW}{dV_{CB'}} \tag{9–37}$$

from Eq. (9–35). To find the value of the capacitance, C, we write

$$C = \frac{dQ}{dV} \tag{9–38}$$

and at the emitter,

$$\begin{aligned} Q &= ep_{e_2} \\ &= \frac{e^2}{kT} p_{n0} V_{eb'} \exp\left(\frac{eV_{EB'}}{kT}\right) \end{aligned} \tag{9–39}$$

from Eq. (9–17). Hence,

$$\begin{aligned} C = \frac{dQ}{dV_{eb'}} &= \frac{e^2}{kT} p_{n0} \exp\left(\frac{eV_{EB'}}{kT}\right) \\ &= \frac{e^2}{kT} p_{e_1} = \frac{eWI_E}{kTD_p} \end{aligned} \tag{9–40}$$

from Eq. (9–33). Thus,

$$\begin{aligned} \mu &= \frac{I_E}{D_p} \frac{dW}{dV_{CB'}} \frac{kTD_p}{eWI_E} \\ &= \frac{kT}{eW} \frac{dW}{dV_{CB'}} \end{aligned} \tag{9–41}$$

For a typical germanium alloy-junction transistor, the fractional change in base width is of the order of two per cent for every volt change between collector and base. At 300° K,

$$\mu = \frac{kT}{eW} \frac{dW}{dV_{CB'}} \sim 0.026\,(0.02) \sim 5 \times 10^{-4} \tag{9–42}$$

9–6 The Hybrid-π Representation

The circuit of Fig. 9–11 shows the transmission line representation of the intrinsic transistor and includes the current generator, I_c, feeding the collector and active base terminals. It is not yet in a practical form because no account has been made of the power which must be drawn from the collector and

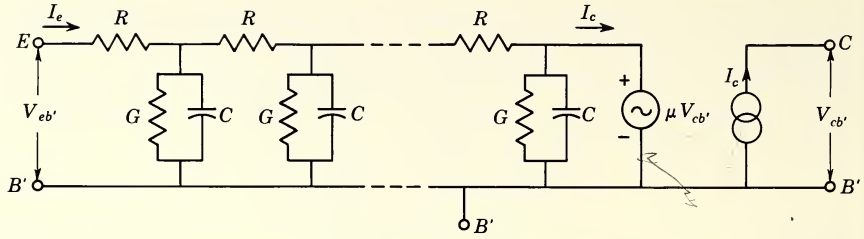


FIG. 9–11. Transmission line analogy showing collector circuit.

active base terminals, C and B', to supply the voltage generator, $\mu V_{cb'}$. However, this circuit can be manipulated into one of several forms which are suitable for circuit analysis. In this section, we will derive the hybrid-π circuit which is used in Chap. 10. (See Sec. 10–1 for a discussion of this circuit.)

The transition between the transmission line representation and the hybrid-π circuit is given in Figs. 9–11 through 9–16. The equivalence of the two circuits can be demonstrated by standard methods of circuit analysis but the process is long and involved. In this account, step-by-step changes will be made in the circuit and approximations will be introduced where desirable.

In the first place, we note that the collector current, I_c, is only slightly smaller than the emitter current, I_e, since $\alpha \sim 1$. Therefore, there is only a small recombination loss of current in the base region. This means that the conductance, G, of Fig. 9–11 is low and its resistance is high. Because of this, it is possible to draw a simple circuit which approximates to the transmission

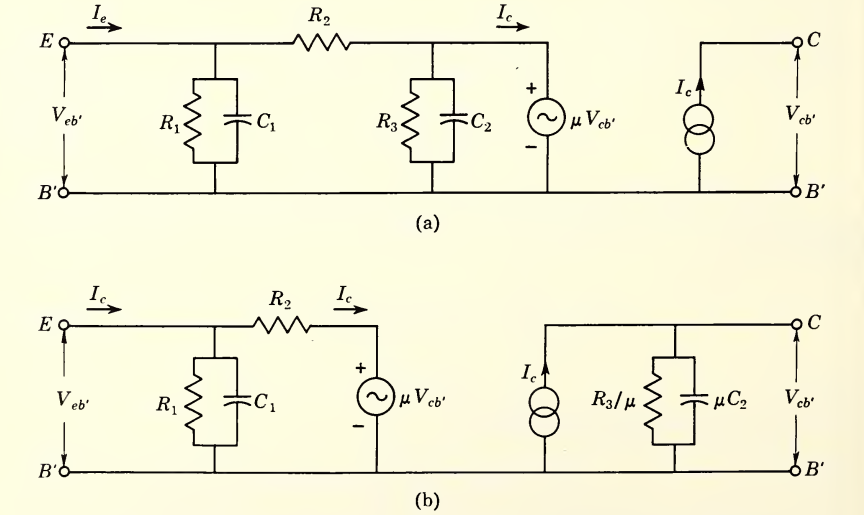


FIG. 9–12. (a) Replacement of transmission line by π equivalent circuit. (b) R_3 and C_2 transferred to output circuit.

line. In Fig. 9–12(a), a π circuit is used in place of the line. R_1 and R_3 now account for the recombination loss of current, and R_2 is the equivalent series resistance of the line. C_1 and C_2 are the capacitance elements across the parallel resistances as before. $R_1 = R_3$ and $C_1 = C_2$ and so the recombination loss and the effective capacitance of the line have been divided into two equal parts; one at each end of the line.

In Fig. 9–12(a), the power to operate the voltage generator, $\mu V_{cb'}$, is obtained from the collector-to-base voltage, $V_{cb'}$. Since R_3 and C_2 are connected in parallel across the constant voltage generator, they have no effect on the input circuit. Thus R_3 and C_2 can be transferred to the output circuit where they become of value R_3/μ and μC_2, respectively, in order to keep the power and phase requirements the same. The circuit now appears as in Fig. 9–12(b).

Before proceeding further, the values of the resistances and the capacitance will be found. When the output is short-circuited to alternating currents ($V_{cb'} = 0$), the collector current, I_c, the emitter current, I_e, and the base current, I_b, are given by the equations,

$$I_c = \alpha I_e \tag{9–43}$$

and

$$I_b = I_e(1 - \alpha) \tag{9–44}$$

Furthermore, at low frequencies where the effects of capacitance can be neglected, the low frequency value of alpha, α_0, can be used and

$$\frac{V_{eb'}}{I_e} = r_\epsilon \tag{9–45}$$

where r_ϵ is the *emitter resistance* which is the incremental resistance of the forward-biased emitter-to-base junction defined in Sec. 4–7. In Fig. 9–12(b), the recombination current flowing through R_1 is the base current, I_b. Hence,

$$R_1 = \frac{V_{eb'}}{I_b} = \frac{V_{eb'}}{I_e(1 - \alpha_0)} = \frac{r_\epsilon}{1 - \alpha_0} \tag{9–46}$$

Also,

$$R_2 = \frac{V_{eb'}}{I_c} = \frac{V_{eb'}}{\alpha_0 I_e} = \frac{r_\epsilon}{\alpha_0} \tag{9–47}$$

Finally, we see from Sec. 5–3F that

$$C_1 = C_2 = C_D \tag{9–48}$$

where C_D is the *diffusion capacitance*. For a base region of constant resistivity, Eq. (5–15) gives the diffusion capacitance of the junction to be

$$C_D = \frac{eW^2}{2kTD_p} I_E \tag{9–49}$$

where I_E is the total emitter current across the junction.

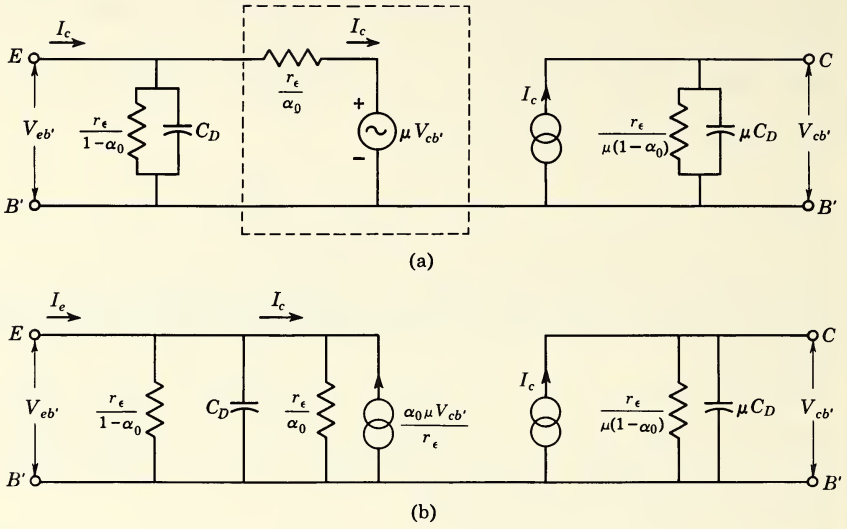


FIG. 9–13. (a) Resistance and capacitance values shown on the circuit of Fig. 9–12(b). (b) Replacement of components in dotted box in part (a) by their Norton equivalent.

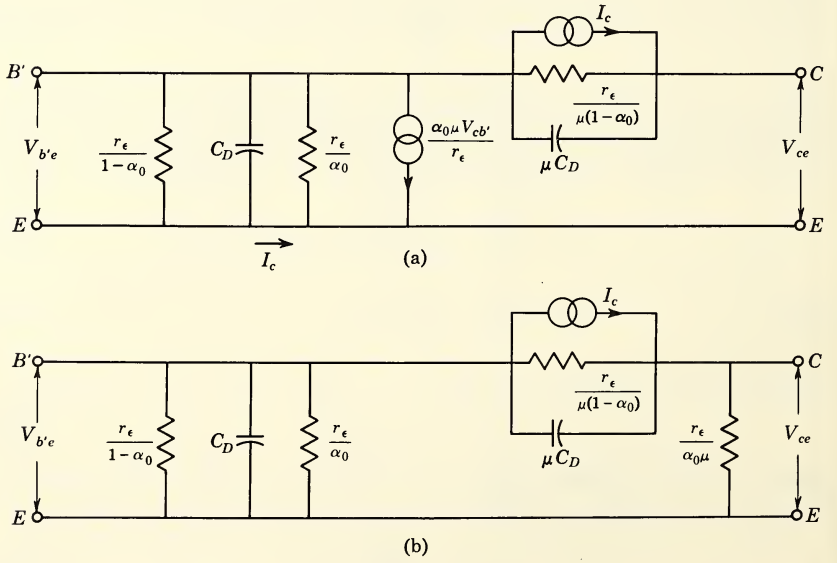


FIG. 9–14. (a) Figure 9–13(b) redrawn in the common emitter connections. (b) Approximation of the current generator in the input circuit.

The values found in the preceding paragraph are shown in Fig. 9–13(a). The voltage generator and the resistance shown in the dotted box in this diagram can be replaced by their Norton equivalent circuit to give Fig. 9–13(b).

Figure 9–13(b) can be redrawn as a common emitter circuit to give Fig. 9–14(a). The input current generator will next be considered and two approximations will be made. In the first place, the current is proportional to the voltage across the CB' terminals. However, for a transistor operating in the common emitter connection, the voltage gain is considerably greater than unity, and so we can write $V_{cb'} \approx V_{ce}$. The value of the current generator now becomes $\alpha_0 \mu V_{ce}/r_\epsilon$ and it will load the collector and emitter terminals as

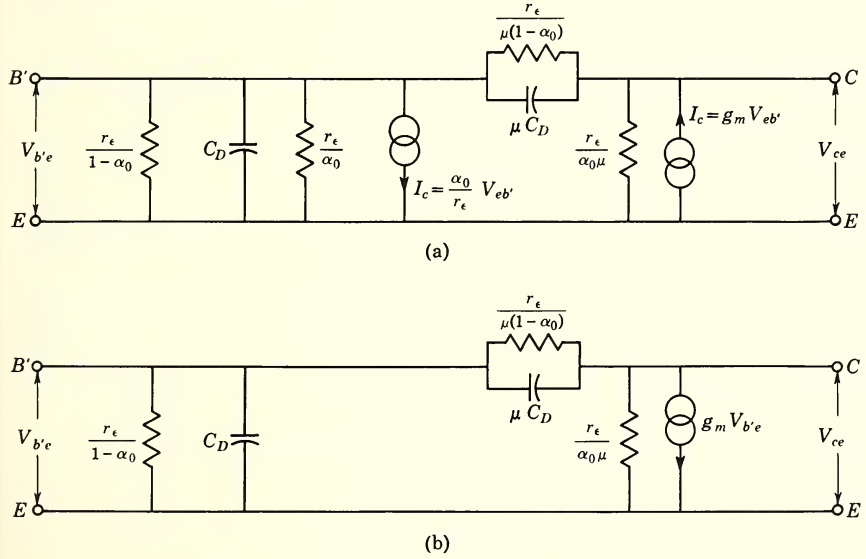


FIG. 9–15. (a) The circuit of Fig. 9–14(b) with the current generator divided into two parts. (b) The π circuit for the intrinsic transistor in the common emitter connection.

if it were a resistance of value $r_\epsilon/(\alpha_0\mu)$ as shown in Fig. 9–14(b). Secondly, we note from Eq. (9–42) that μ is very small. In the input circuit, therefore, the generator current can be neglected in comparison with I_c, which is flowing in a connecting lead. The two approximations are shown in Fig. 9–14(b).

The remaining current generator in Fig. 9–14(b) has been divided in Fig. 9–15(a) into two generators which have a common junction at the emitter. The two circuits are electrically identical since a constant current generator may be split in this manner without altering its effect on the circuit. From Eq. (9–47),

$$I_c = \frac{\alpha_0}{r_\epsilon} V_{eb'} = g_m V_{eb'} \tag{9–50}$$

where $g_m = \alpha_0/r$. Thus the generator now in the input circuit supplies a current which is proportional to the voltage across its terminals. This generator can be replaced by a resistance, and, because of the direction of current flow, the resistance is negative and has the value

$$-\frac{V_{eb'}}{g_m V_{eb'}} = -\frac{1}{g_m} = -\frac{r_\epsilon}{\alpha_0} \tag{9-51}$$

This negative resistance cancels out the existing positive resistance of the same magnitude already across the input terminals. The final circuit for the intrinsic transistor is given in Fig. 9–15(b). It is a π representation of the approximated transmission line for a common emitter circuit. The direction

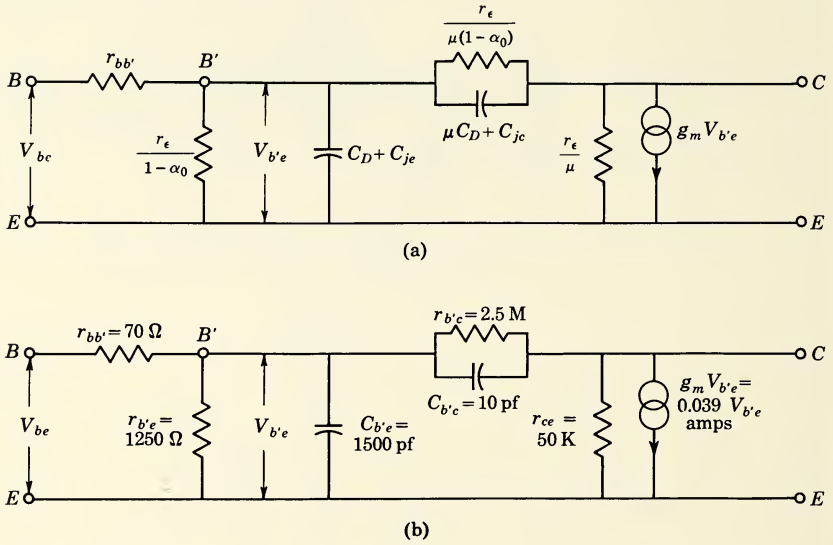


FIG. 9–16. (a) The hybrid-π circuit for the common emitter connection. (b) Typical values for the hybrid-π circuit.

of the current generator and the polarity of the voltage to which it is related have both been changed to conform to current practice.

The complete hybrid-π circuit is shown in Fig. 9–16(a). Here, the three important components outside of the intrinsic transistor, namely $r_{bb'}$, C_{je}, and C_{jc}, have been added. (Compare this common emitter circuit with the common base circuit given in Fig. 9–3.) In addition, α_0 has been replaced by unity where this approximation causes little error. Figure 9–16(b) shows the common designations for the various elements of the circuit. Values given in this diagram apply to a typical medium frequency, alloy-junction, *pnp* transistor operating at $I_E = 1$ milliamp and $V_{CE} = -5$ volts.

The quantity, $g_m = \alpha_0/r$, has the dimensions of mhos or amps per volt. It is sometimes called the "intrinsic transconductance." From Sec. 4–7,

$r_e = 25.8/I_E$ ohms, where I_E is in milliamps. Thus g_m is known for a given emitter current and, in the typical case shown in Fig. 9–16(b), $g_m = 39$ milliamps per volt when $I_E = 1$ milliamp.

The elements of the hybrid-π circuit are substantially independent of frequency in the usable range of the transistor. In addition, the variation in the values of the components of the circuit as the quiescent voltage or current are changed can readily be calculated. These advantages are further discussed in Chap. 10.

9–7 Variation of Alpha with Emitter Current

The value of alpha for the majority of transistors lies in the range 0.95 to 0.995. The CE short-circuit current gain, β, is given by $\beta = \alpha/(1 - \alpha)$ and so

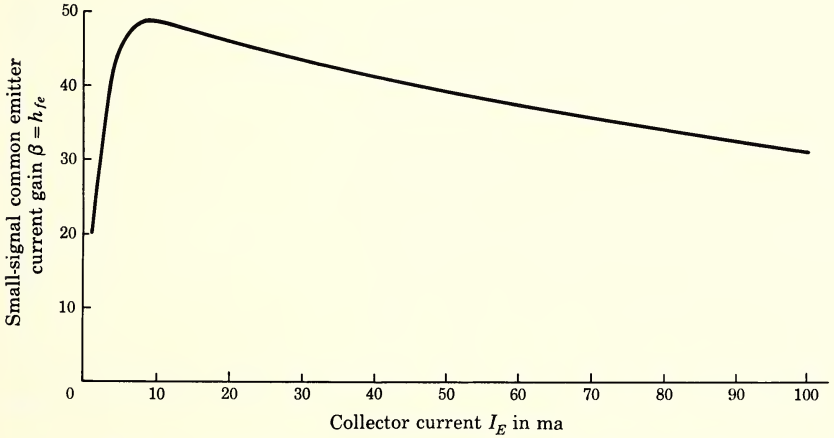


FIG. 9–17. Variation in h_{fe} with emitter current for a *pnp* transistor. Considerable variation in this curve is found for different transistors.

the corresponding range of values of β is 19 to 199. More specifically, since

$$\beta = \frac{\alpha}{1 - \alpha}$$

$$\frac{d\beta}{d\alpha} = \frac{1}{(1 - \alpha)^2} = \frac{\beta + 1}{1 - \alpha}$$

or

$$\frac{d\beta}{\beta + 1} = \frac{d\alpha}{1 - \alpha} \tag{9–52}$$

Since alpha is close to unity, Eq. (9–52) shows that the CE current gain is extremely sensitive to small changes in alpha. In this section, the variation of alpha with quiescent emitter current is discussed.

There are three mechanisms which account for the change in alpha.

1. At very low current densities: recombination in the emitter-base depletion region causes a reduction in current entering the base.
2. At high current densities: the number of majority carriers in the base rises, and an electric field is produced which aids the passage of minority charges through the base region. This results in an increase in alpha.
3. At very high current densities: the large number of carriers in the base increases the base conductivity and decreases both the emitter efficiency and the lifetime of the carriers in the base region. This causes a drop in alpha as the emitter current is increased (see Fig. 9–17).

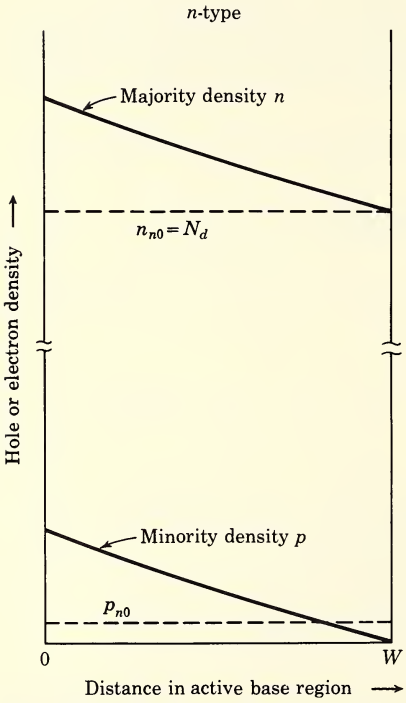


FIG. 9–18. Carrier density in the base region when the emitter current density is high.

At low current densities, it is found that recombination in the depletion regions of the emitter-to-base junction forms a high proportion of the total emitter current. This is responsible for a small value of alpha when the emitter current is low. As the emitter current is increased, recombination processes capture a smaller proportion of the emitter charge carriers and alpha rises.

Up to now, it has been assumed that small-current conditions applied to the transistor and that the majority carrier density was substantially constant with distance. When the emitter current is high, this condition cannot be accepted and the variation of the majority carrier density with distance in the base region must be considered. Figure 9–18 shows the majority and minority carrier density in the base region under large-current operation. This carrier density profile is obtained by the condition for charge neutrality in the base region

$$n + N_a = p + N_d \tag{9–53}$$

and, assuming $N_a \sim 0$,

$$n - p = N_d \tag{9–54}$$

Under small-current conditions, p is assumed to be small everywhere, giving $n = N_d$ (a constant), as shown previously. Under large-current operation,

the value of p at $x = 0$ no longer can be neglected in comparison with N_a, and so Eq. (9–54) applies as shown in Fig. 9–18.

Charge neutrality is not exactly obeyed in the base region. The conditions shown in Fig. 9–18 lead to a diffusion current formed by electrons moving across the base region from left to right because of the concentration gradient. Since the emitter efficiency is close to unity, the net electron current is approximately zero and some mechanism must be operating to cause an equal current in the opposite direction. This mechanism is that of drift of electrons in an electric field. The field is produced by the diffusion of electrons which produces a small imbalance of the charge neutrality condition and produces a field $\mathscr{E}$. The difference between the positive and negative charge density distributions is very small, and Eq. (9–53) can still be used in the analysis except where the electric field is to be calculated.

Since the total electron current is zero and both drift and diffusion take place,

$$J^- = en\mu_n\mathscr{E} + eD_n\frac{dn}{dx} = 0$$

or

$$\mathscr{E} = -\frac{D_n}{n\mu_n}\frac{dn}{dx} \tag{9–55}$$

Since dn/dx is negative, the field is from left to right. Assuming that the departure from charge neutrality is small, we may differentiate Eq. (9–54) with respect to x and write

$$\frac{dn}{dx} = \frac{dp}{dx} \tag{9–56}$$

From the Einstein equation of Sec. 4–8,

$$\frac{D_n}{\mu_n} = \frac{D_p}{\mu_p} \tag{9–57}$$

and so

$$\mathscr{E} = -\frac{D_p}{n\mu_p}\frac{dp}{dx} \tag{9–58}$$

This electric field also affects the hole current which is flowing, since

$$\begin{aligned} J^+ &= ep\mu_p\mathscr{E} - eD_p\frac{dp}{dx} \\ &= -\frac{ep}{n}D_p\frac{dp}{dx} - eD_p\frac{dp}{dx} \\ &= -eD_p\frac{dp}{dx}\left(1 + \frac{p}{n}\right) \end{aligned} \tag{9–59}$$

Under small-current conditions, $p \ll n$ and the current flow is wholly by diffusion. As p increases, part of the total current is carried by drift and, from

Eq. (9–59), this can be considered as an apparent increase in the diffusion constant, D_p. In the limit, $p \to n$, and D_p apparently doubles. The field is responsible for aiding the holes in their passage across the base and less recombination takes place since they spend less time in the n-type base region. This effect is responsible for an increase in alpha as the emitter current is raised.

In Sec. 5–3B, the emitter efficiency was given as

$$\gamma = 1 - \frac{\sigma_b}{\sigma_e}\frac{W}{L_n} \tag{9–60}$$

where σ_b and σ_e are the conductivities of the base and emitter regions respectively. Under large-current conditions, there are many charge carriers present in the base region as shown in Fig. 9–18. Thus the base conductivity, σ_b, increases and the emitter efficiency falls. The lifetime for holes in the base region, τ_p, is also decreased by the presence of a large number of electrons and this reduces the transport factor, β_0^*. As the emitter current increases to very large values, these effects take control, and alpha decreases.

9–8 The Drift Transistor

So far in this book, the transistor has been assumed to have a base region of constant resistivity. Under small-current conditions, it has been shown that the injected minority carriers move across the base by diffusion and that no appreciable electric field exists in the base. For this reason, we may use the term "diffusion transistor" for the device although recognizing, from Sec. 9–7, that an electric field will be present when the emitter current density is high.

In the diffusion transistor, the transit time for minority carriers across the base forms an upper limit on the frequency of operation of the device. Reducing the thickness of the base decreases the transit time and increases the usable frequency range of operation, but if the base is made too thin, other limitations arise. For example, the base spreading resistance increases as the base thickness is reduced and, whatever changes in thickness, conductivity, and area of the regions are made, a point is reached where it is uneconomical to seek a higher frequency of operation for the diffusion transistor.

The drift transistor overcomes some of the limitations of the diffusion transistor and can be operated at frequencies up to a few thousand megacycles per second. Devices are commercially available which give substantial gain in a common emitter circuit at several hundred megacycles per second. This is accomplished by reducing the transit time of the minority carriers in the base region and by ensuring that the collector capacitance is small.

The drift transistor is manufactured with a nonuniform base conductivity. As shown below, this results in a "built-in" electric field which aids the

minority carrier motion across the base region. Both drift and diffusion occur in the drift transistor, and the combination of the two current-carrying mechanisms reduces the transit time and so increases the upper frequency response. In a *pnp* transistor of this type, the *n*-type impurity is diffused into the base region under very carefully controlled conditions of temperature and pressure so that an impurity gradient is produced in the base region. In many cases, the impurity density in the base can be assumed to drop approximately exponentially from the emitter to the collector junction and this results in a constant "built-in" field. There are many methods of producing a drift transistor; all of them involve the diffusion of an *n*- or *p*-type impurity from a vapor into the base region of a transistor. The manufacturers' literature should be consulted for the method used for any particular type (see also Sec. 5–6).

The way in which the impurity density varies with distance in a typical drift transistor is shown in Fig. 9–19. Close to the collector junction, the base resistivity is high and the material is near-intrinsic. The charge depletion region at the collector end of the junction extends well into the base as shown, and this results in a low collector-to-base junction capacitance [Eq. (3–24)]. This is an important characteristic of the drift transistor and is partly responsible for its good high frequency performance.

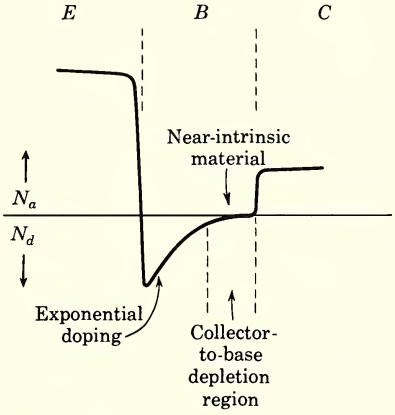


FIG. 9–19. Structure of drift transistor.

The electric field which is present by virtue of a nonuniform base conductivity can be calculated by considering the majority charge distribution in the base. For a *pnp* drift transistor, the majority carrier density has the profile shown in Fig. 9–19, and so electrons will diffuse in the direction of decreasing concentration, that is, from left to right. As previously discussed in Sec. 9–7, this slightly unbalances the condition of charge neutrality and results in an electric field, $\mathscr{E}$, which produces a drift of electrons in the opposite direction. Writing the equations for the total current density of electrons in the base region, we see that

$$J^- = e\mu_n n\mathscr{E} + eD_n \frac{dn}{dx}$$

$$= 0 \tag{9–61}$$

since no net majority carrier flow occurs in the base. Thus,

$$\mathscr{E} = -\frac{D_n}{\mu_n n}\frac{dn}{dx} \tag{9–62}$$

From the Einstein relation of Sec. 4–8,

$$\frac{D_n}{\mu_n} = \frac{kT}{e} \tag{9–63}$$

and so

$$\mathscr{E} = -\frac{kT}{en}\frac{dn}{dx} \tag{9–64}$$

If $\mathscr{E}$ is to be made constant throughout the base, Eq. (9–64) can be integrated to give

$$n = n_0 \exp\left(-\frac{e}{kT}\mathscr{E}x\right) \tag{9–65}$$

where $n = n_0$ at $x = 0$. Furthermore, if $n = n_W$ at $x = W$, we can solve Eq. (9–65) for $\mathscr{E}$ to give

$$\mathscr{E} = \frac{kT}{eW}\log_e\left(\frac{n_0}{n_W}\right) \tag{9–66}$$

This is the constant "built-in" field that is produced by the exponential majority carrier density distribution. It has a positive value since $n_0 > n_W$ and is acting in the direction from emitter to collector. For the majority carriers (electrons), as we have seen, the currents attributable to drift and diffusion are equal and opposite and the net current is zero. For minority carriers (positive holes), the drift and diffusion currents add and the total hole current density is

$$J^+ = e\mu_p p\mathscr{E} - eD_p\frac{dp}{dx} \tag{9–67}$$

where dp/dx is a negative quantity. The total hole current density across the base, J^+, is a constant. Therefore, the first order differential equation (9–67) can be solved to give

$$p = \frac{J^+}{e\mathscr{E}\mu_p}\left\{1 - \exp\left[(x - W)\right]\frac{\mathscr{E}\mu_p}{D_p}\right\} \tag{9–68}$$

where the constant of integration has been evaluated by noting that p = 0 at $x = W$. From Eqs. (9–64) and (9–66)

$$\frac{\mathscr{E}\mu_p}{D_p} = \frac{1}{W}\log_e\left(\frac{n_0}{n_W}\right) \tag{9–69}$$

Thus,

$$p = \frac{J^+}{e\mathscr{E}\mu_p}\left\{1 - \exp\left[\left(\frac{x - W}{W}\right)\log_e\left(\frac{n_0}{n_W}\right)\right]\right\} \tag{9–70}$$

At the beginning of the base region, where $x \ll W$,

$$p = \frac{J^+}{e\mathscr{E}\mu_p}\left\{1 - \frac{n_W}{n_0}\right\} \approx \frac{J^+}{e\mathscr{E}\mu_p} \tag{9–71}$$

if $n_W/n_0 \ll 1$. In practice, the doping is such that n_W/n_0 is of the order of 1/8. Thus, nearly 90 per cent of the current is carried by drift at the beginning of the base region but as $x \to W$, diffusion increases. The transit time of the holes across the base is much reduced by the presence of the "built-in" field, and so the transit time is reduced.

The term "diffused-base transistor" has come into general use in the semiconductor device industry. The process of diffusion, which is now standard in the production of high frequency units, results in a base region with a nonuniform impurity density (approximately exponential). The analysis of the drift transistor applies to some extent to all diffused-base transistor constructions. Whenever the base conductivity is nonuniform, an internal electric field will be developed in the base region. Diffused-base transistors are now designed for particular applications. By varying the base width, conductivity values, junction areas, etc., the important properties of the transistor which determine its high frequency and pulse characteristics can be suited to the application (see Chap. 10).

9–9 Transistor Noise

The random noise generated in a device determines its ultimate sensitivity as an amplifier. There are three sources of noise in a transistor: *resistance noise* caused by *thermal agitation*; *shot noise* caused by the *discrete nature* of the charge carriers; and *modulation* or *semiconductor noise*.

Resistance noise is due to the random motion of the holes and electrons in a material and can be represented by a mean square noise voltage of the

$$\overline{v^2} = 4kTR\,\Delta f \tag{9–72}$$

form where k is Boltzmann's constant, T is in °K, R is the resistance of the material and Δf is the band of frequencies over which the noise is measured. This is known as "white" noise since it has a constant frequency spectrum, that is, for a constant bandwidth, Δf, $\overline{v^2}$ does not depend on frequency.

Shot noise is produced wherever a current is flowing in the transistor, and it also is "white" noise. The current is carried by individual carriers (holes or electrons) and so it fluctuates slightly about its mean value. The mean square fluctuation in the current is

$$\overline{i^2} = 2eI\Delta f \tag{9–73}$$

where I is the average current. If this mean square current is flowing through a resistance, R, it will produce a mean square voltage across the terminals of

$$\overline{v^2} = 2eIR^2\,\Delta f \tag{9–74}$$

Modulation or semiconductor noise is also known as *flicker noise*. It is entirely a low frequency phenomenon and is of considerable importance in

low frequency and d-c amplifiers. Its origin is thought to lie in the semiconductor crystal imperfections and in surface effects. It is not "white" noise, but can be represented by a mean square voltage of the form

$$\overline{v^2} = \frac{AI^2\,\Delta f}{f} \tag{9–75}$$

where A is a constant and I is the current flowing. Thus, the mean square voltage is inversely proportional to the frequency, f. This type of noise is predominant below about 1 kc/sec in most transistors.

The transistor adds a certain amount of noise power to any signal that it amplifies. However, the input signal to the transistor includes thermal noise of its own, generated by the source resistance, R_S. An assessment of the

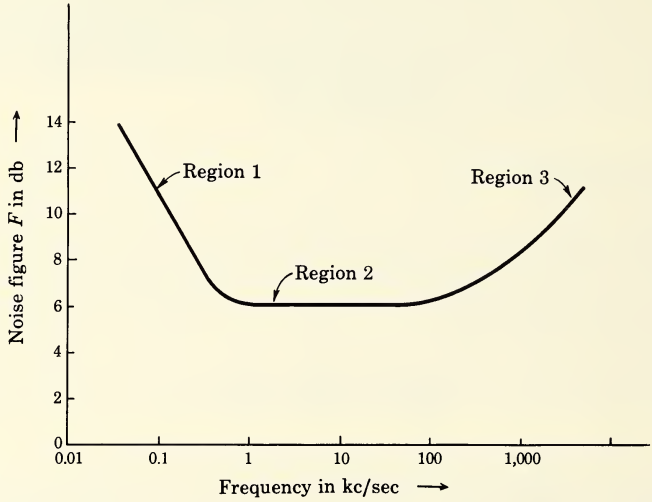


FIG. 9–20. Noise figure of a transistor as a function of frequency. Source resistance, $R_S = 500$ ohms.

transistor's noise performance can be obtained by finding the fractional increase in noise caused by the transistor. The *noise figure*, F, is defined to be the ratio of the total noise output power to the noise output power attributable solely to the thermal noise in the source. Hence,

$$F = \frac{A_p P_{SN} + P_{TN}}{A_p P_{SN}} = 1 + \frac{P_{TN}}{AP_{SN}} \tag{9–76}$$

where P_{TN} is the internally generated transistor noise power, P_{SN} is the thermal noise in the signal, and A_p is the transistor power gain. The value of P_{SN} per unit bandwidth does not vary with frequency as shown by Eq. (9–72), assuming a resistive source. The power gain, A_p, and the transistor noise, P_{TN}, are frequency dependent, and thus the noise figure of the transistor will vary with frequency.

The noise figure is usually quoted in decibels where

$$F_{db} = 10 \log_{10} \left(1 + \frac{P_{TN}}{A_p P_{SN}}\right) \tag{9–77}$$

Figure 9–20 shows a curve of the noise figure versus frequency for a typical transistor. Between about 1 kc/sec and 100 kc/sec, the noise figure is 6 db. From Eq. (9–77), this means that the total noise power at the output terminals, under these conditions, is composed of one part determined by the external noise source (R_S) and three parts generated in the transistor. Noise figures as low as 1 db can be obtained using selected transistors under ideal conditions.

The three regions of Fig. 9–20 can be identified from Eqs. (9–72) through (9–77). In region 1, the modulation noise is high in the transistor, and the noise figure increases as the frequency decreases. In region 2, modulation noise can be ignored, and thermal and shot noise predominate. The noise figure increases in region 3 mainly because the power gain of the transistor decreases as the frequency is raised. The demarcation frequency between regions 2 and 3 is approximately the geometrical mean of the alpha and beta cutoff frequencies. The "spot" noise figure given by many manufacturers is measured at 1 kc/sec.

Figure 9–20 has been drawn for a source resistance of 500 ohms. For most transistors, the noise figure is found to be at its lowest value when R_S is between 100 and 1000 ohms. For low noise operation, the transistor should be operated with emitter current and collector-to-base voltage in the low-to-medium range.

9–10 Surface Effects in Transistors

Transistors are very sensitive to the conditions existing on the semiconductor surfaces, and surface contamination in device manufacture is a major problem. In the early days of transistors, surface contamination caused a slow deterioration of the transistor characteristics. This problem has been largely overcome by extensive surface treatment and elaborate encapsulation of the device, but the surface properties still play some part in determining the characteristics of the transistor.

If the surface of a semiconducting crystal is rough, there will be many recombination centers present and the lifetimes of the charge carriers will be reduced (see Sec. 1–20). Since the lifetimes in the bulk material are unchanged, the hole and electron densities close to the surface are reduced and there will be a flow of carriers towards the surface. The net current flow is zero since both types of carrier are involved. Conditions at the surface are characterized by the *surface recombination velocity*, s, defined by

$$s = \frac{J_s}{e(p_s - p_0)}$$

where p_s and p_0 are the thermal equilibrium hole densities at the surface and in the bulk material respectively, and J_s is the hole current density flowing to

the surface. Values of s vary from about 2000 cm sec^{-1} (for a sand-blasted surface) to less than 200 cm sec^{-1} (for an electrolytically etched surface). In the base region of a *pnp* transistor, any positive holes which travel to the surface and recombine are lost from the collector current. This loss may be reduced by making the surface recombination velocity small and by ensuring that the collector subtends a large solid angle to every point on the emitter. The diffused-base transistor is good in this respect since the drift field aids the minority carriers to pass directly from the emitter to the collector.

Many substances which are present in the atmosphere, e.g., water, form a surface layer on germanium or silicon which is only a few molecules thick. This layer readily becomes charged and produces a high electric field at the surface of the semiconductor and alters the conditions in the device. Manufacturing processes are now being used which "passivate" the surface to reduce the effects of surface films and so prevent the variation of transistor characteristics with time.

The collector cutoff current, I_{CBO}, is markedly dependent on the surface conditions of the semiconducting material. The wide variation in I_{CBO} between transistors with apparently the same construction is often the result of differences in surface conditions.

9–11 Punch-through

In Secs. 5–3C and 9–5 base width modulation in a diffusion transistor was discussed. It was shown that a change in collector-to-base voltage, V_{CB}, is responsible for a change in the active width of the base region, W. The width of the depletion region at the collector end of the base is proportional to V_{CB}^{n}, where n is approximately $-\frac{1}{2}$ for the alloy-junction diffusion transistor, and $-\frac{1}{3}$ for the diffused-base drift transistor. When a high reverse collector-to-base voltage is applied to the transistor, the depletion width may be so large that the active base region disappears entirely and the emitter-to-base and collector-to-base depletion regions merge. Transistor action is no longer possible under these conditions. This phenomenon is known as *punch-through* and the particular collector-to-base voltage at which it occurs is the *punch-through voltage*.

Punch-through is one of the conditions which determine the upper limit of allowable collector-to-base voltage. This voltage will be specified by the manufacturer. Punch-through is not necessarily harmful to the transistor. However, if the resistance in the external circuit permits an excessively high current to flow under punch-through conditions, the transistor will be damaged.

9–12 Epitaxial Films

Starting in 1960, several manufacturers added the *epitaxial* process to the production of diffused-base transistors. In transistor parlance the term epitaxial

refers to the deposition of a thin, highly resistive layer of single crystal semiconducting material on a low resistivity substrate where the crystal orientation of the upper layer is the same as that of the substrate. The principal use of epitaxial films has been to improve the collector saturation voltage and speed of diffused-base transistors.

The epitaxial film is deposited by a vapor process. The semiconductor material in the form of a compound is vaporized and a chemical or thermal decomposition is made to occur in the vicinity of the single crystal substrate. The substrate is heated and semiconductor atoms from the vapor deposit epitaxially on the substrate surface. By adding compounds of doping elements to the vapor stream, deposition of the doping element also occurs, and an epitaxial film of the desired resistivity and thickness is grown on the substrate.

In the conventional mesa transistor (see Fig. 5–15), the collector material is of fairly high resistivity to avoid collector-to-base breakdown at low voltages. However, this means that the ohmic resistance of the collector

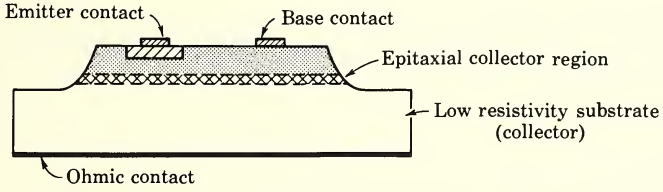


FIG. 9–21. Epitaxial mesa diffused-base structure. Compare with Fig. 5–15.

region is appreciable and the collector-to-emitter saturation voltage is correspondingly high. In addition, the high resistivity material has a long lifetime and minority charges will, therefore, be stored in the collector region for a comparatively long time and over a large distance. This causes a limit to be set on the switching time of the transistor.

The epitaxial mesa transistor is shown in Fig. 9–21. It employs a collector wafer of low resistivity, on top of which is deposited the epitaxial high resistance layer as shown in the diagram. The reduced resistance of the collector region ensures that the collector-to-emitter saturation voltage is small and it may be as low as one tenth of its value for the conventional structure (e.g., down to 0.4 volts for $I_c = 100$ ma). The lifetime of minority carriers in the high resistivity epitaxial layer is much smaller and so switching times are reduced. Other advantages are lower collector-to-emitter capacitance and the retention of high current gain at high current levels.

BIBLIOGRAPHY

Cote, Alfred J., Jr., and J. Barry Oakes, *Linear Vacuum-tube and Transistor Circuits*, New York: McGraw-Hill Book Company, Inc., 1961

DeWitt, David, and Arthur L. Rossoff, *Transistor Electronics*, New York: McGraw-Hill Book Company, Inc., 1957

Gärtner, Wolfgang W., *Transistors: Principles, Design and Applications*, Princeton, N.J.: D. Van Nostrand Company, Inc., 1960

Middlebrook, R. D., *An Introduction to Junction Transistor Theory*, New York: John Wiley & Sons, Inc., 1957

Pettit, Joseph Mayo, and Malcolm Myers McWhorter, *Electronic Amplifier Circuits, Theory and Design*, New York: McGraw-Hill Book Company, Inc., 1961

Wolfendale, E., ed., *The Junction Transistor and its Applications*, London: Heywood & Company Ltd., 1958

PROBLEMS

9–1 Explain the meaning of the base transport factor, β^*. Find β^* as a function of frequency and calculate the frequency at which β^* drops to 90 per cent of its low frequency value if $W/L_p = 0.0785$ and $\tau_p = 60$ μsec.

ans. 870 kc/sec

9–2 Given the hybrid-π circuit of Fig. 9–16(b), find the value of the four h parameters.

9–3 Explain in words, equations, and diagrams the phenomenon of base width modulation.

9–4 A *pnp* germanium transistor has abrupt junctions and is operated at room temperature. The following details of the transistor are known. Emitter: $\rho_p = 0.07$ ohm cm; Base: $\rho_n = 1.6$ ohm cm; Collector: $\rho_p = 0.40$ ohm cm; Metallurgical base width $= 2 \times 10^{-3}$ cm; $\alpha_0 = 0.98$ (assumed equal to β_0^*); Collector-to-base area $= 5 \times 10^{-3}$ cm^2; $V_{CB} = 10$ volts (reverse bias); $I_E = 1$ ma. Determine the following quantities, (a) contact potential between collector and base, (b) collector-to-base depletion width, (c) $r_{b'e}$, (d) μ from Eq. (9–42).

ans. 0.268 volts, 4.16×10^{-4} cm, 1290 ohms, 3.25×10^{-4}

9–5 For the transistor of Prob. 9–4, calculate (a) $C_{b'e}$ with the condition $C_D \gg C_{je}$, (b) $r_{b'c}$, (c) $C_{b'c}$ taking μC_D into account, (d) r_{ce}, (e) g_m. Draw the hybrid-π circuit for the transistor showing the numerical values for all the components.

10

High Frequency and Pulse Operation of the Transistor

The hybrid-π equivalent circuit, which was derived in Chap. 9, is used here to determine the upper frequency response of the transistor stage. The analysis is limited to the common emitter configuration and includes a calculation of the gain and bandwidth for an iterative stage. The hybrid-π model is then compared with the actual transistor, and it is shown that a simple, empirical modification of the circuit leads to a more exact representation at high frequencies.

The hybrid-π circuit is also used initially in the analysis of pulsed operation of the transistor. However, the concept of charge control is briefly developed and it is shown that this new approach offers advantages in analyzing and understanding the transient conditions of switching.

10–1 The Hybrid-π Circuit

Figure 10–1 shows the hybrid-π model of the transistor that will be used in the first part of this chapter. It is an active π circuit with the base spreading

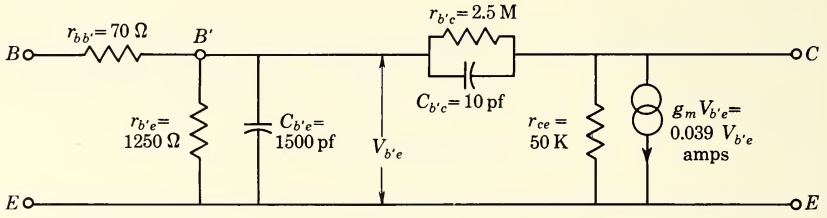


FIG. 10–1. The hybrid-π circuit with typical values. (I_E = 1 ma).

resistance, $r_{bb'}$, added to the input lead. This circuit is developed from the transmission-line analog of the transistor in Sec. 9–6.

The important elements of this circuit are identified below.

$r_{bb'}$, the *base spreading resistance*, is the ohmic resistance between the base lead and the active base region. It is substantially independent of operating conditions.

$r_{b'e}$ is the a-c resistance between the active base, B', and the emitter terminal, E. Its value is given by $r_\epsilon/(1 - \alpha_0)$ where $r_\epsilon = kT/eI_E = 25.8/I_E$ mhos at 300°K where I_E is in milliamps. In the region where α_0 can be considered a constant, $r_{b'e}$ is inversely proportional to I_E.

$C_{b'e}$ is the sum of the emitter *junction capacitance*, C_{je}, and the emitter-to-base *diffusion capacitance*, C_D. In many cases C_D predominates. From Eq. (5–15), C_D is proportional to I_E.

$C_{b'c}$ is the sum of the collector *junction capacitance*, C_{jc}, and a small fraction of the diffusion capacitance, C_D. C_{jc} usually predominates and is proportional to V_{CB}^n where n is between $-\frac{1}{2}$ and $-\frac{1}{3}$ depending on the type of transistor (see Sec. 3–4).

g_m is the *transconductance* and is of magnitude α_0/r_ϵ. It is, therefore, approximately proportional to I_E and has the value of 0.039 mhos when I_E = 1 ma at 300°K.

All of these values should be considered approximate and considerable deviations will be found in practice. To the first order of accuracy, all the

elements are independent of frequency, and the change in the major components of the circuit can be calculated as I_E or V_{CE} is changed. These two factors provide major advantages over other transistor representations which might be used.

An analysis of the CE stage at high frequencies will be carried out using the hybrid-π circuit. This gives an adequate representation for most transistors at frequencies an order of magnitude less than the alpha-cutoff frequency. For higher frequencies, modifications must be made to the circuit if it is to be used with accuracy. These modifications are briefly discussed in Sec. 10–6.

10–2 The Miller Effect

Figure 10–2(a) shows the a-c circuit of a transistor amplifier in the CE connection with a load R_L and a constant current input source. Figure 10–2(b)

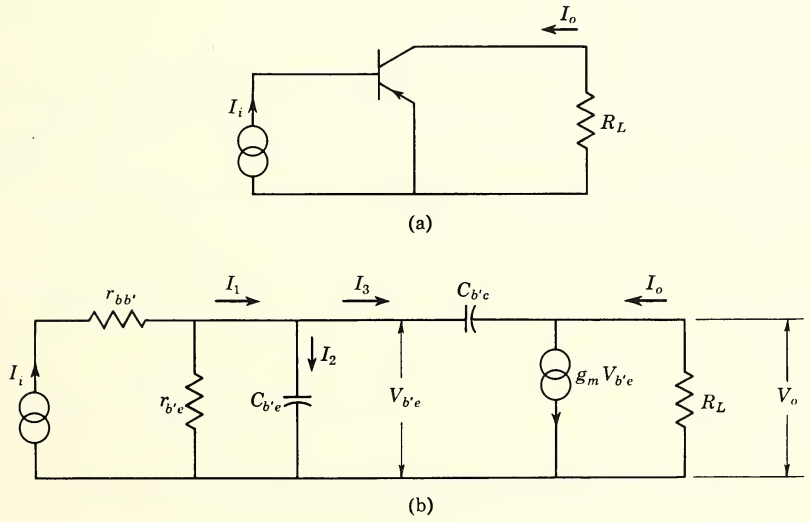


FIG. 10–2. (a) The a-c circuit of a CE amplifier. (b) Simplified hybrid-π equivalent.

shows a simplified hybrid-π equivalent. Two components have been omitted from the complete hybrid-π circuit of Fig. 10–1. These are: r_{ce}, which is significantly larger than the load resistance, and $r_{b'c}$, which is usually so high that its effect can be neglected. Typical values for these components are given in Fig. 10–1.

This section will consider the simplified hybrid-π circuit and show that the effective capacitance between B' and E is increased by the presence of $C_{b'c}$, and that a new equivalent circuit can be drawn (Fig. 10–3). This is analogous to the *Miller effect* in vacuum tube circuitry.

From Fig. 10–2(b),

$$I_1 = I_2 + I_3 \tag{10–1}$$

and

$$I_2 = j\omega C_{b'e} V_{b'e} \tag{10–2}$$

Furthermore,

$$I_3 = (V_{b'e} - V_o) j\omega C_{b'c} \tag{10–3}$$

In practice, $I_3 \ll I_o$ and so

$$V_o = -g_m V_{b'e} R_L \tag{10–4}$$

Therefore,

$$\begin{aligned} I_1 &= j\omega C_{b'e} V_{b'e} + j\omega C_{b'c}(V_{b'e} - V_o) \\ &= j\omega V_{b'e}[C_{b'e} + (1 + g_m R_L) C_{b'c}] \end{aligned} \tag{10–5}$$

and

$$\frac{I_1}{V_{b'e}} = j\omega[C_{b'e} + (1 + g_m R_L) C_{b'c}] \tag{10–6}$$

Thus, as far as the input side of the transistor is concerned, the effect of the feedback current, I_3, is to increase the capacitance to the value C, where

$$C = C_{b'e} + (1 + g_m R_L) C_{b'c} \tag{10–7}$$

This effect was first noted by J. M. Miller for the case of vacuum tubes in 1919.

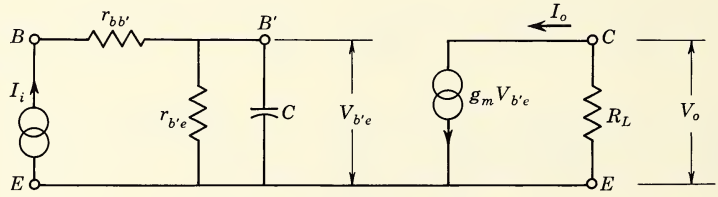


FIG. 10–3. Simplified equivalent circuit showing effective capacitance, C.

When the collector load, R_L, is so small that the output circuit is unaffected by the feedback current, I_3, the equivalent circuit can be drawn as in Fig. 10–3. This is the usual case for an iterative resistance-capacitance coupled stage.

Using the values given in Sec. 10–1 and Fig. 10–1, with an emitter direct current of 1 ma, and with $R_L = 1000$ ohms,

$$\begin{aligned} C &= 1500 + (1 + 39)10 \\ &= 1900 \text{ pf} \end{aligned} \tag{10–8}$$

This capacitance is in parallel with $r_{b'e}$ and the two elements form a CR circuit which will be shown to limit the frequency response of the transistor stage.

10-3 Frequency Response of the Transistor in the Common Emitter Connection: Hybrid-π Model

From Fig. 10-3, the impedance of the parallel resistance-capacitance combination, $r_{b'e}C$. at an angular frequency ω is

$$Z = \frac{r_{b'e}}{1 + j\omega r_{b'e}C} = \frac{V_{b'e}}{I_i} \tag{10-9}$$

The current gain of the transistor in the common emitter connection is

$$A_i = \frac{I_o}{I_i} = g_m V_{b'e} \frac{Z}{V_{b'e}} = \frac{g_m r_{b'e}}{1 + j\omega r_{b'e}C} \tag{10-10}$$

from Eq. (10-9). Now, $g_m = \alpha_0/r_e$, and $r_{b'e} = r_e/(1 - \alpha_0)$, hence,

$$A_i = \frac{\alpha_0}{1 - \alpha_0} \frac{1}{1 + j\omega r_{b'e}C} = \frac{\beta_0}{1 + j\omega r_{b'e}C} \tag{10-11}$$

At low frequencies, Eq. (10-11) shows that $A_i = \beta_0$. Recalling that

$$C = C_{b'e} + (1 + g_m R_L)C_{b'c}$$

Eq. (10-11) gives the frequency response of the transistor in the CE connection, that is,

$$A_i = \frac{\beta_0}{1 + j\omega r_{b'e}[C_{b'e} + (1 + g_m R_L)C_{b'c}]} \tag{10-12}$$

Under short-circuit conditions with $R_L \to 0$,

$$A_i = \frac{\beta_0}{1 + j\omega r_{b'e}(C_{b'e} + C_{b'c})} \tag{10-13}$$

At an angular frequency, ω_β, where

$$\omega_\beta = 2\pi f_\beta = \frac{1}{r_{b'e}(C_{b'e} + C_{b'c})} \approx \frac{1}{r_{b'e}C_{b'e}} \tag{10-14}$$

the current gain of the stage will have fallen to $1/\sqrt{2}$ of its low frequency value. The short-circuit current gain as a function of angular frequency can be written as

$$\beta(\omega) = \frac{\beta_0}{1 + j(\omega/\omega_\beta)} = \frac{\beta_0}{1 + j(f/f_\beta)} \tag{10-15}$$

Figure 10-4 shows a plot of the magnitude of the short-circuit current gain versus frequency for a hybrid-π circuit. The current gain in the figure is given in decibels (db) where the number of db $= 20 \log_{10} (I_o/I_i)$. At low frequencies, $\beta_0 = 50$ corresponds to 34 db of gain. At the frequency f_β, the gain has fallen to $\beta_0/\sqrt{2}$ which corresponds to 31 db or a loss of 3 db. The frequency where there is a 3 db loss of gain is a measure of the useful

band-width of the transistor when used in a stage with a low frequency gain of β_0. It will be shown in the next section that this does not limit the operation of the common emitter stage to frequencies lower than f_β, provided that a lower value of gain at low frequencies can be tolerated. f_β is known as the *beta-cutoff frequency* of the transistor.

At frequencies considerably in excess of f_β, the imaginary term in Eq. (10–15) becomes much greater than unity. The drop in gain for one octave increase in frequency is then 6 db, and the curve shown in Fig. 10–4 becomes a

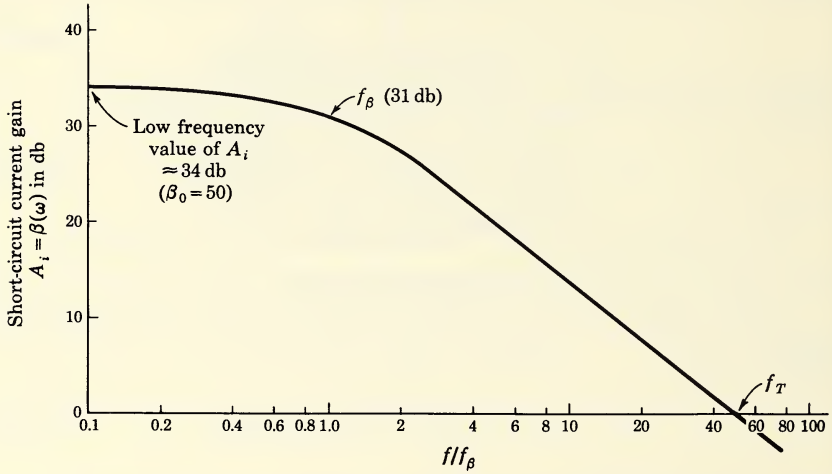


FIG. 10–4. High frequency response of a typical hybrid-π circuit showing f_β and f_T. β_0 is assumed to be 50.

straight line. The frequency where $|A_i|$ is unity is named f_T, and can be obtained from Eq. (10–13) by noting that $\omega_T r_{b'e}(C_{b'e} + C_{b'c}) \gg 1$. Thus,

$$\omega_T = 2\pi f_T = \frac{\beta_0}{r_{b'e}(C_{b'e} + C_{b'c})} \approx \frac{\beta_0}{r_{b'e}C_{b'e}} \tag{10–16}$$

From Eqs. (10–14) and (10–16),

$$\beta_0 f_\beta = f_T \tag{10–17}$$

for the hybrid-π circuit, and so f_T is the *gain-bandwidth product* for the transistor in the CE connection. f_T is also a useful parameter in pulse operation of the transistor and it will be discussed further in Secs. 10–6 and 10–10.

The calculations shown above, starting with Eq. (10–13), apply only when the output terminals are short-circuited to alternating currents. If a high collector load is inserted so that the capacitance C [Eq. (10–7)], includes an appreciable contribution due to the term $g_m R_L C_{b'c}$, then the angular frequency corresponding to a loss in gain of 3 db will occur at

$$\omega_{3\text{ db}} = \frac{1}{r_{b'e}[C_{b'e} + (1 + g_m R_L)C_{b'c}]}$$

from Eq. (10–12). Assuming $g_m R_L \gg 1$,

$$\omega_{3\,\mathrm{db}} = \frac{1}{r_{b'e}C_{b'e} + r_{b'e}g_m R_L C_{b'c}} = \frac{1}{r_{b'e}C_{b'e} + \beta_0 C_{b'c} R_L}$$

when the values for $r_{b'e}$ and g_m are substituted from Sec. 10–1. The 3 db frequency is, therefore, shown to include a time constant term, $\beta_0 C_{b'c} R_L$, in the denominator. $C_{b'c}$ is the measured output capacitance in the CB configuration and is commonly designated C_{ob} in manufacturers' data sheets. The output capacitance in the CE connection (sometimes designated C_{oe}) is $\beta_0 C_{ob}$ and the resultant output time constant $C_{oe}R_L = \beta_0 C_{b'c} R_L$ appears in the denominator of the equation. To obtain maximum bandwidth from a transistor amplifier, R_L is chosen so that $\beta_0 C_{b'c} R_L \ll r_{b'e}C_{b'e}$. Note that reducing R_L slightly increases the current gain of a stage but materially reduces the voltage gain. In a multistage amplifier, it is the current gain which is of importance and R_L is usually small enough that the inequality is satisfied.

10–4 Frequency Response of the Iterative Common Emitter Stage: Hybrid-π Model

Figure 10–5 shows the equivalent circuit of a transistor stage where the resistance R_S shunts the input current generator. In a multistage transistor

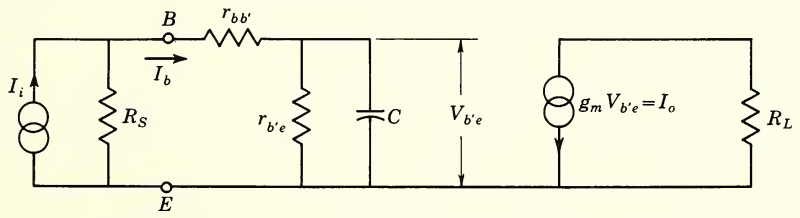


FIG. 10–5. Equivalent circuit of typical CE iterative stage.

amplifier, R_S will be composed of the load resistance of the previous stage in parallel with resistances in the interstage coupling network. The presence of R_S has an effect on the frequency response of the stage as shown below.

If I_b is the alternating current entering the base terminal of the transistor,

$$\frac{I_b}{I_i} = \frac{R_S}{Z + r_{bb'} + R_S} \tag{10–18}$$

where Z is given in Eq. (10–9). The voltage across Z is $V_{b'e}$, hence,

$$\frac{V_{b'e}}{I_i} = Z\frac{I_b}{I_i} = \frac{ZR_S}{Z + r_{bb'} + R_S} \tag{10–19}$$

Substituting for Z from Eq. (10–9), and putting $R_t = R_S + r_{b'e} + r_{bb'}$, we have

$$\frac{V_{b'e}}{I_i} = \frac{R_S[r_{b'e}/(1 + j\omega r_{b'e}C)]}{r_{b'e}/(1 + j\omega r_{b'e}C) + r_{bb'} + R_S}$$

$$= \frac{R_S r_{b'e}}{R_t + j\omega r_{b'e}C(R_S + r_{bb'})} \tag{10–20}$$

The current gain of the stage is

$$A_i = \frac{I_o}{I_i} = \frac{g_m V_{b'e}}{I_i} = \frac{g_m R_S r_{b'e}}{R_t + j\omega r_{b'e} C(R_S + r_{bb'})} \tag{10–21}$$

The current gain of the stage is 3 db down from its low frequency value at an angular frequency, $\omega_{3\,\mathrm{db}}$, given by

$$\omega_{3\,\mathrm{db}} = \frac{R_t}{R_S + r_{bb'}} \frac{1}{r_{b'e} C} = \frac{R_t}{R_S + r_{bb'}} \omega_\beta \tag{10–22}$$

The low frequency gain from Eq. (10–21) is

$$A_{io} = \frac{g_m R_S r_{b'e}}{R_t} \tag{10–23}$$

The product of the low frequency gain and the angular frequency is given by

$$A_{io}\omega_{3\,\mathrm{db}} = \frac{g_m R_S r_{b'e}}{R_t} \frac{R_t}{R_S + r_{bb'}} \frac{1}{r_{b'e} C}$$

$$= \frac{R_S}{R_S + r_{bb'}} \frac{g_m}{C} \tag{10–24}$$

If $R_S \gg r_{bb'}$, $A_{io}\omega_{3\,\mathrm{db}} = g_m/C$ and this value can be used as a *figure of merit* for the transistor.

A consideration of Eqs. (10–22) and (10–23) shows that the transistor can be used at frequencies in excess of the beta-cutoff frequency provided that a lower current gain can be tolerated. For the transistor of Fig. 10–1,

$$f_\beta = \frac{\omega_\beta}{2\pi} = \frac{1}{2\pi r_{b'e} C} = \frac{10^{12}}{2\pi(1250)1900} = 670 \text{ kc/sec}$$

If $R_S = \infty$, the low frequency gain, $A_{io} = g_m r_{b'e} = 49$ and $\omega_{3\,\mathrm{db}} = \omega_\beta$. When $R_S = 1000$ ohms,

$$A_{io} = \frac{0.039(1000)1250}{2320} = 21$$

and

$$f_{3\,\mathrm{db}} = \frac{\omega_{3\,\mathrm{db}}}{2\pi} = \frac{2320}{1070} 670{,}000 = 1.44 \text{ Mc/sec}$$

If R_S is further reduced, the stage gain decreases and the 3 db frequency increases. Note that the stage gain includes the shunting effect of R_S.

10–5 Frequency Response of the Transistor in the Common Base Connection: Simple Model

The calculations which have been made on the CE hybrid-π representation of the transistor suggest a simple means of comparing the frequency response

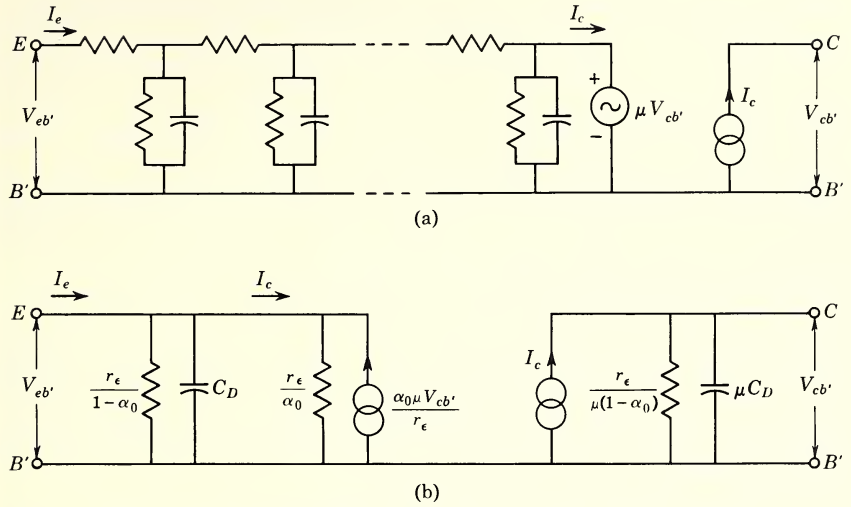


FIG. 10-6. (a) Transmission line analog of base region. (b) A simplified representation of the transmission line analog.

of the CB and CE connections. In Fig. 10-6(a) we repeat the transmission line analog of the base region of the intrinsic transistor previously given in Fig. 9-11, and Fig. 10-6(b) shows the simplified representation given before in Fig. 9-13(b). This circuit can be used to find the frequency response of the transistor in the CB connection.

Figure 10-7 has been drawn for the intrinsic transistor operated under short-circuit output conditions. When the emitter-to-base junction capacitance is included, the input capacitance, $C_{b'e}$, is given by

$$C_{b'e} = C_D + C_{je} \tag{10-25}$$

where C_{je} is the junction capacitance from emitter to base. The combination of the two resistances, $r_\epsilon/(1 - \alpha_0)$ and r_ϵ/α_0, in parallel is r_ϵ, and so

$$I_i \frac{r_\epsilon}{1 + j\omega C_{b'e} r_\epsilon} = I_c \frac{r_\epsilon}{\alpha_0}$$

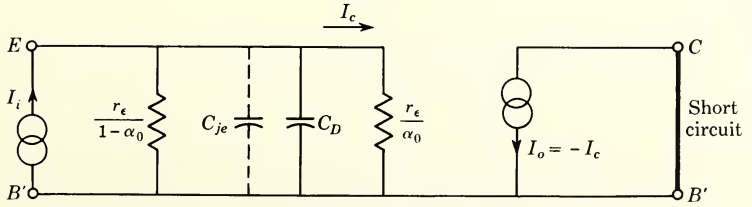


FIG. 10-7. The intrinsic transistor operating in the common emitter connection with short circuited output. Emitter-to-base junction capacitance shown with dashed line.

and the current gain in the short-circuited common emitter connection is

$$A_i = \alpha(\omega) = \frac{I_c}{I_i} = \frac{\alpha_0}{1 + j\omega C_{b'e} r_\epsilon} \tag{10–26}$$

The *alpha-cutoff frequency*, f_α, is defined as the frequency where A_i has dropped to $1/\sqrt{2}$ of its low frequency value. Thus,

$$\omega_\alpha = 2\pi f_\alpha = \frac{1}{C_{b'e} r_\epsilon} \tag{10–27}$$

from Eq. (10–26) and that equation can be written as

$$\alpha(\omega) = \frac{\alpha_0}{1 + j(\omega/\omega_\alpha)} \tag{10–28}$$

By comparison with the value for the beta-cutoff angular frequency [Eq. (10–14)], and noting that $r_{b'e} = r_\epsilon/(1 - \alpha_0)$,

$$\omega_\beta = \frac{1 - \alpha_0}{r_\epsilon C_{b'e}} \approx \frac{\omega_\alpha}{\beta_0} \tag{10–29}$$

For the hybrid-π circuit, therefore,

$$f_\alpha \approx f_T \approx \beta_0 f_\beta \tag{10–30}$$

10–6 Comparison of the Hybrid–π Circuit with the Actual Transistor

The hybrid-π circuit is a simplified representation of the transistor which serves as a starting point for electronic circuit design. When the transistor is operated at high frequencies or under pulse conditions, however, the hybrid-π circuit provides only an approximate model for the analysis of transistor behavior. Fortunately, this model can be improved with some simple modifications, and it is then adequate for most types of circuit design.

When measurements of transistor parameters are attempted, the experimenter becomes aware of a considerable spread in the quantities that he is trying to measure. Variations in materials and manufacturing techniques are responsible for most of these deviations, and the circuit designer must account for the resulting spread in the characteristics. Thus, electronic circuit components are seldom specified to an accuracy of better than 10 per cent, and the designer endeavors to produce a circuit which will operate even when all the parameters and conditions take their most unfavorable values. This is known as the "worst-case" design philosophy and is important in the design of large-scale electronic equipment where the failure of any circuit must be avoided. It is for this reason that circuit electronics is not an exact subject and, consequently, a model of a solid state

or vacuum device is usually adequate if it represents the device to within an accuracy of about 10 per cent within the frequency band considered.

The alpha-cutoff angular frequency has been given by Eq. (10–27) as

$$\omega_\alpha = \frac{1}{C_{b'e} r_\epsilon} \approx \frac{1}{C_D r_\epsilon}$$

We can substitute for C_D and r_ϵ from Eqs. (5–15) and (4–45), to give

$$\omega_\alpha = \frac{2kTD_p}{eW^2 I_E} \frac{eI_E}{kT} = \frac{2D_p}{W^2} \tag{10–31}$$

The hybrid-π circuit was derived by simplifying the transmission line equivalent to the transistor. When the a-c continuity equation (9–21) is solved, the exact relationship for the diffusion transistor is

$$\omega_\alpha = 2.43 \frac{D_p}{W^2} \tag{10–32}$$

Therefore, the hybrid-π circuit shows an error of about 20 per cent in its prediction of the alpha-cutoff frequency from the physical constants of the intrinsic (diffusion) transistor.

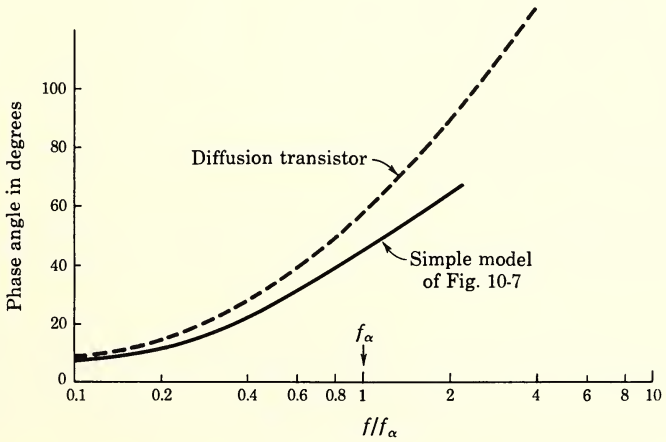


FIG. 10–8. Phase difference between output and input current in the common base connection as a function of f/f_α for a diffusion transistor and for the model of Fig. 10–7.

The measured phase between the output and the input current for a diffusion transistor also varies considerably from that predicted by the simple model. This is shown in Fig. 10–8 where the diffusion transistor and the hybrid-π model are compared. The difference in phase angle between the two curves is approximately proportional to f/f_α. Empirically, we may account

for this phase difference by an additional term in the short-circuit current gain [Eq. (10–28)], to give

$$\alpha(\omega) = \frac{\alpha_0 \exp(-jm\omega/\omega_\alpha)}{1 + j(\omega/\omega_\alpha)} \tag{10–33}$$

where m has a value of about 0.2 for a diffusion transistor. The additional term has a magnitude of unity so the magnitude of the current generator is

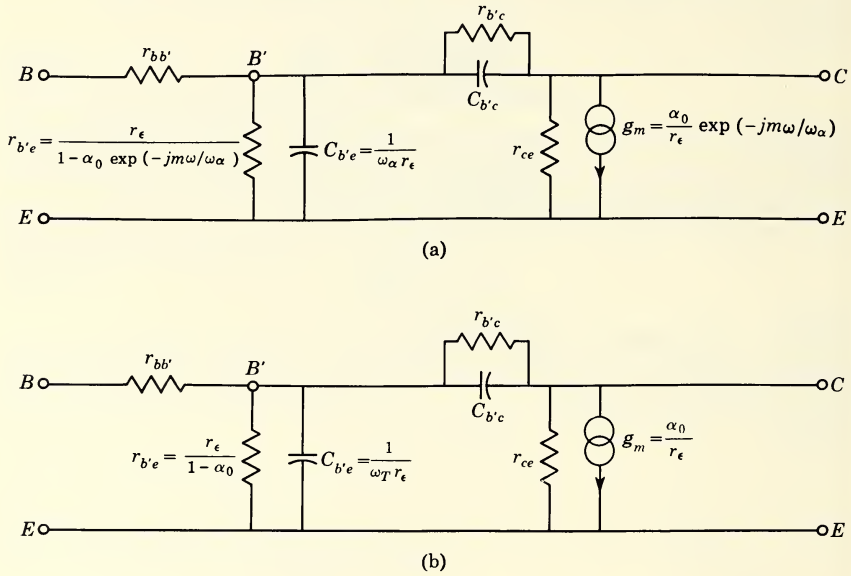


FIG. 10–9. (a) Phase-modified hybrid-π circuit based on Eq. (10–33). (b) Hybrid-π circuit based on ω_T.

unaltered, and only its phase is changed. For a drift or diffused-base transistor, the value of m is between 0.2 and 1 according to the value of the aiding field.

A phase-modified hybrid-π circuit based on Eq. (10–33) is shown in Fig. 10–9(a). In this circuit, the extra phase difference is considered to be contained in the α_0 term so that wherever α_0 appears, $\alpha_0 \exp(-jm\omega/\omega_\alpha)$ must be substituted. Thus, the current generator has the value of

$$g_m = \frac{\alpha_0}{r_\epsilon} \exp(-jm\omega/\omega_\alpha) \tag{10–34}$$

and $r_{b'e}$ is given by

$$r_{b'e} = \frac{r_\epsilon}{1 - \alpha_0 \exp(-jm\omega/\omega_\alpha)} \tag{10–35}$$

From Eq. (10–26), the alpha-cutoff frequency can be calculated.

$$A_i = \frac{\alpha_0 \exp(-jm\omega/\omega_\alpha)}{1 + j\omega C_{b'e} r_\epsilon} \tag{10–36}$$

Since the magnitude of the numerator is unchanged,

$$\omega_\alpha = \frac{1}{C_{b'e} r_\epsilon}$$

as before and $C_{b'e}$ can be chosen to conform with the measured value of ω_α. The current gain of this modified hybrid-π circuit corresponds to the empirical formula, Eq. (10–33). In this way, we can retain the hybrid-π circuit and also account for the difference in phase between the output and input current. The value of m is chosen to be appropriate to the transistor in use. In the analysis that follows we will consider only the diffusion transistor and put $m = 0.2$.

For the common emitter stage, the value of f_T is important. This is the frequency at which the magnitude of the CE current gain drops to unity. From Eqs. (10–11), (10–35), and (10–27),

$$A_i = \frac{\alpha_0 \exp(-jm\omega/\omega_\alpha)}{1 - \alpha_0 \exp(-jm\omega/\omega_\alpha)} \frac{1}{1 + j \dfrac{\omega}{\omega_\alpha[1 - \alpha_0 \exp(-jm\omega/\omega_\alpha)]}}$$

$$= \frac{\alpha_0 \exp(-jm\omega/\omega_\alpha)}{1 - \alpha_0 \exp(-jm\omega/\omega_\alpha) + j(\omega/\omega_\alpha)} \tag{10–37}$$

when $|A_i| = 1$, $\omega = \omega_T$ and so

$$\left|1 - \alpha_0 \exp\left(\frac{-jm\omega_T}{\omega_\alpha}\right) + j\left(\frac{\omega_T}{\omega_\alpha}\right)\right| = \left|\alpha_0 \exp\left(\frac{(-jm\omega_T}{\omega_\alpha}\right)\right| \tag{10–38}$$

From the hybrid-π circuit we know that $\omega_T \sim \omega_\alpha$ and, since $m = 0.2$, the exponentials can be replaced by the first two terms of a series expansion. Thus,

$$\left|1 - \alpha_0 + j0.2\alpha_0\left(\frac{\omega_T}{\omega_\alpha}\right) + j\left(\frac{\omega_T}{\omega_\alpha}\right)\right| = \left|\alpha_0 - j\,0.2\alpha_0\left(\frac{\omega_T}{\omega_\alpha}\right)\right|$$

or

$$\left|1 - \alpha_0 + j(1 + 0.2\alpha_0)\left(\frac{\omega_T}{\omega_\alpha}\right)\right| = \left|a_0 - j\,0.2\alpha_0\left(\frac{\omega_T}{\omega_\alpha}\right)\right|$$

When $\alpha_0 \to 1$, $1 - \alpha_0 \to 0$, $1 + 0.2\alpha_0 \to 1.2$, and $|\alpha_0 - j\,0.2\alpha_0(\omega_T/\omega_\alpha)| \approx 1$. Hence,

$$\omega_T \approx \frac{\omega_\alpha}{1.2} \tag{10–39}$$

From Eqs. (10–39) and (10–32), therefore,

$$\omega_T \approx \frac{\omega_\alpha}{1.2} \approx \frac{2D_p}{W^2} \tag{10–40}$$

The more simple circuit of Fig. 10–9(b) is often used for high frequency and pulse analysis. The elements g_m and $r_{b'e}$ do not include a phase term and

the capacitance $C_{b'e}$ has a value which is determined from the measured value of ω_T, according to the equation

$$C_{b'e} = C_{je} + C_D \approx C_D = \frac{1}{\omega_T r_e} \tag{10–41}$$

It is, therefore, a simple hybrid-π circuit based on the measured parameter ω_T. In practice, ω_T will not be measured directly but will be inferred from measurements of gain in the CE connection at high frequencies. f_T is then K times the frequency where the gain is K, and where K is less than 10. This assumes a 6 db per octave loss in gain with frequency.

In the previous equations, we have neglected C_{je} in comparison with C_D. When the hybrid-π circuit is used under small-signal, class A, amplifying conditions with a large emitter current, this is a justifiable assumption. If the emitter current is low, however, the emitter-to-base voltage is small and C_{je} can no longer be ignored. In this case, as the emitter current is decreased, ω_T will fall.

10–7 Saturation

Saturation of a transistor can be explained by reference to Fig. 10–10. In part (a) of the figure, a *pnp* transistor is operating with a collector supply,

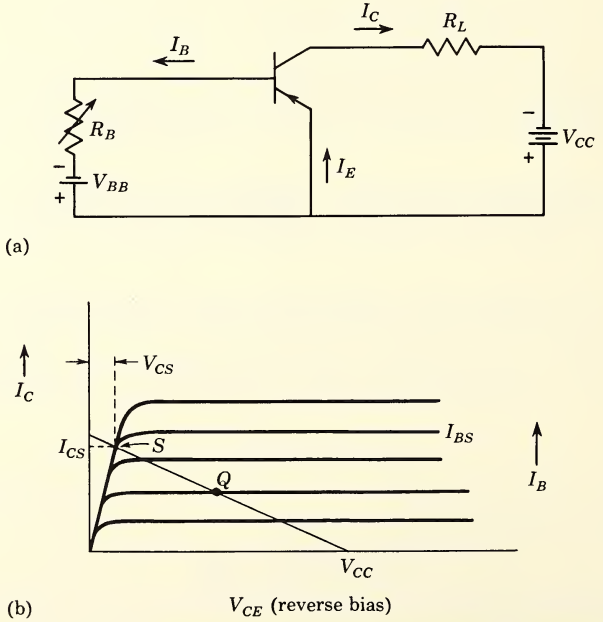


FIG. 10–10. (a) Operating circuit for a *pnp* transistor. (b) Load line showing saturation conditions.

V_{CC}, and a collector load resistance, R_L. The output characteristics in part (b) show the load line corresponding to the circuit. All currents and voltages shown are d-c values.

When the transistor is operating normally, V_{CE} is negative and the collector-to-base junction is reverse biased. As the base current is increased, the operating point moves up the load line from, say, Q to S. *Saturation* is said to occur at S where an increase in I_B causes no significant change in I_C. Now, using the sign convention of Chap. 8,

$$I_E = I_C + I_B \tag{10–42}$$

and

$$V_{CE} = V_{CC} + I_C R_L \tag{10–43}$$

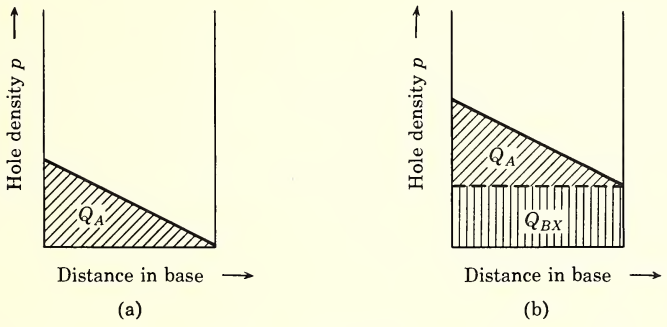


FIG. 10–11. (a) Variation of hole density with distance in the base region of a *pnp* diffusion transistor operating just below saturation. (b) The same transistor in saturation.

The value of V_{CE} at saturation is called V_{CS} (see Fig. 10–10), and it is usually a fraction of a volt. The critical value of I_C when saturation occurs is I_{CS}, where

$$I_{CS} = \frac{V_{CE} - V_{CC}}{R_L} \approx \frac{-V_{CC}}{R_L} \tag{10–44}$$

from Eq. (10–43). Neglecting I_{CEO}, we can write,

$$I_C = h_{FE} I_B \tag{10–45}$$

which applies only when I_B is at, or below, the saturation value, I_{BS}. Thus the base current at which the transistor goes into saturation is

$$I_{BS} = \frac{I_{CS}}{h_{FE}} \approx \frac{-V_{CC}}{h_{FE} R_L} \tag{10–46}$$

Saturation is usually avoided in amplifiers because it produces distortion. In pulse operation of the transistor, it may sometimes be desirable to operate in saturation when the device is switched "on" since the power which has to be supplied to keep the transistor in this condition is low.

Conditions in the base region under normal transistor operation (just below saturation) are shown in Fig. 10–11(a). The hole density decreases across the base region and falls to approximately zero at the collector-base junction at the right of the diagram. In Fig. 10–11(b), the transistor is shown in saturation. The saturation collector current, I_{CS}, is proportional to the slope of the hole density curve at the collector-base junction as before, since current is still transferred across the base region. However, this current is limited to I_{CS} under saturation conditions, whereas I_B and I_E are above the values necessary for saturation to occur. Thus more charge is entering the base region via the emitter lead than is being removed via the collector lead. The charge stored in the device rises to the value shown in Fig. 10–11(b) until the recombination current has increased sufficiently to give the necessary balance between I_{CS}, I_E, and I_B.

The horizontal dashed line drawn on Fig. 10–11(b) divides the stored charge into two parts. Q_A is approximately the same as in part (a) of the figure if the effects of increased recombination are neglected. Q_{BX} is the extra stored charge in the base region of the transistor under saturation conditions. When the saturated transistor is switched off, this charge must be removed from the base region before the collector current drops below its saturated value. The time for this to take place is known as the *storage time* and is discussed in Secs. 10–9 and 10–11.

10–8 Switching the Transistor on: Use of the Hybrid-π Model

As a starting point in the discussion of the pulse operation of the transistor, consider the circuit of Fig. 10–12(a). Assume that the emitter-to-base junction is initially biased with a reverse voltage so that no emitter current is flowing in the transistor. At $t = 0$, the base current assumes a value I_{b_1}, as shown in Fig. 10–12(b). Saturation does not occur in this circuit because $R_L = 0$. Part (c) of the figure gives the hybrid-π circuit which was derived for small-signal operation of the transistor at high frequences. This model will be used to predict how I_c varies with time. Since we are using a small-signal model under large-signal transient conditions, we cannot expect a high accuracy. In this section, therefore, we shall be content to derive the form of the variations in output current, and so prepare for the charge control method of analysis given in Sec. 10–10.

In order to switch the transistor on, the direct voltage across the base-emitter junction must reach a sufficiently high forward value to cause current to flow. This is about 0.2 volts for a germanium transistor and 0.6 volts for a silicon device. The capacitance in the input circuit, in the general case, has previously been given as

$$C = C_{b'e} + (1 + g_m R_L) C_{b'c} \qquad (10\text{–}7)$$

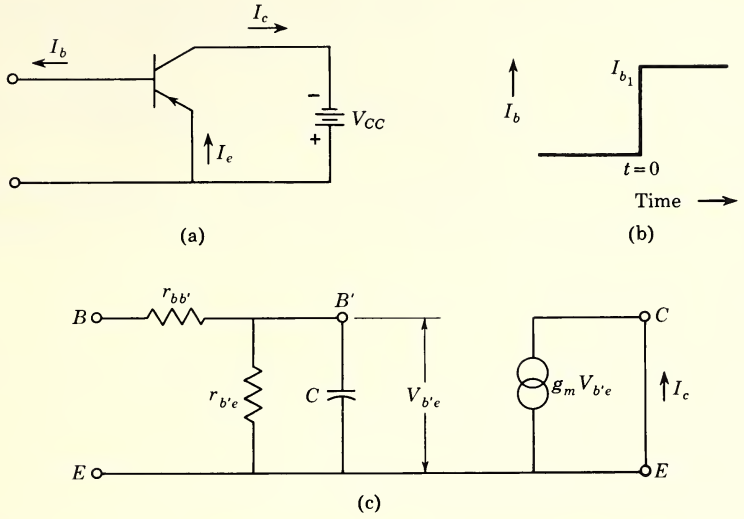


FIG. 10–12. (a) *pnp* transistor circuit. (b) Input current waveform. (c) The simplified hybrid-π circuit.

where

$$C_{b'e} = C_{je} + C_D \tag{10–47}$$

During the time required to achieve forward conduction, however,

$$g_m = C_D = 0$$

because $I_E = 0$. Therefore

$$C = C_{b'e} + C_{b'c} = C_{je} + C_{jc} \approx C_{je} \tag{10–48}$$

C_{je} will usually dominate Eq. (10–48) because the voltage across the emitter-to-base junction is very low (see Sec. 3–4). The resistance, $r_{b'e}$, becomes infinite because $I_E = 0$, and so the input circuit reduces to $r_{bb'}$ in series with C_{je} as in Fig. 10–13.

The capacitance, C_{je}, is determined by the emitter-to-base voltage and if V_R is the reverse voltage on the emitter-to-base junction corresponding to a capacitance C_{je1}, the charge initially on this capacitance is $V_R C_{je1}$. When the transistor is operating with a forward voltage $V_{b'e}$ and its input capacitance is C_{je2}, the charge on the capacitance (of the opposite sign) is $V_{b'e}C_{je2}$. Since the flow of current in the base lead, I_{b_1}, is responsible for this change in charge, the delay time before collector current flows can be found from

$$I_{b_1} t = V_{b'e}C_{je2} + V_R C_{je1} \tag{10–49}$$

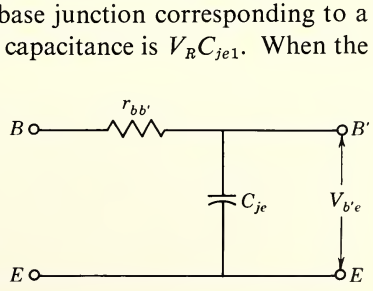


FIG. 10–13. Input circuit of the transistor before initiation.

For a small time delay, both V_R and the emitter-to-base junction capacitance should be small. The variation in C_{je} with voltage will be specified by the manufacturers for a particular transistor. When I_{b_1} is supplied from a voltage source with a finite source resistance, R_S, the product of C_{je} and $R_S + r_{bb'}$ is the time constant of the circuit which represents the transistor while it is still in the "off" state.

After emitter current starts to flow in the transistor, the simplified hybrid-π circuit of Fig. 10–12(c) can be applied. When the base current, I_{b_1}, is constant,

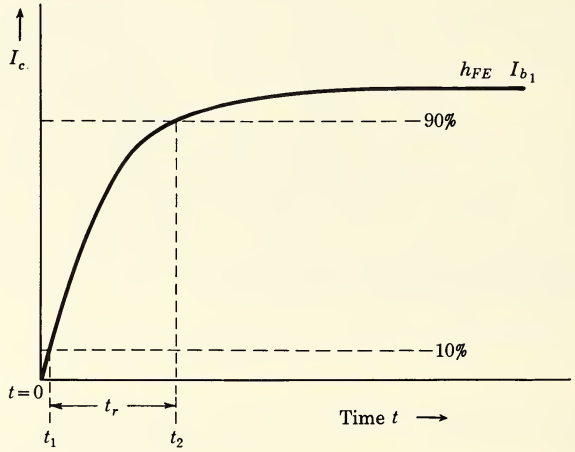


FIG. 10–14. Rise of collective current with time (h_{FE} and ω_T assumed constant).

$V_{b'e}$ rises to a final value of $I_{b_1}r_{b'e}$ with a time constant of $r_{b'e}C$. From Eqs. (10–7) and (10–47) we see that C is a function of emitter current and hence of time. The collector current at any time is given by

$$I_c = g_m V_{b'e} = g_m I_{b_1} r_{b'e} \left[1 - \exp\left(-\frac{t}{r_{b'e}C}\right)\right]$$

$$\approx h_{FE} I_{b_1} \left[1 - \exp\left(-\frac{t\omega_T}{h_{FE}}\right)\right] \qquad (10\text{–}50)$$

from Eqs. (10–11) and (10–16). Since we are dealing with large-signal operation, β, which was used in the previous sections, has been replaced by h_{FE}. In Eq. (10–50), the time constant, h_{FE}/ω_T, varies as the current rises, since both C and h_{FE} are functions of I_E. This variation is considered in more detail in Sec. 10–10 and, in the remainder of this section, h_{FE} and ω_T will be considered constant so that we may derive the form of the rise in output current.

Figure 10–14 shows the collector current as a function of time, starting from the instant that the collector current starts to flow and ignoring the time delay before initiation. From the figure, t_1 and t_2 are the times when the

collector current reaches 10 per cent and 90 per cent of its final value. The *rise time*, t_r, is defined to be

$$t_r = t_2 - t_1 \tag{10–51}$$

From Eq. (10–50), we see that the final collector current is $h_{FE}I_{b_1}$, and so

$$0.1 = 1 - \exp\left(\frac{-t_1\omega_T}{h_{FE}}\right) \tag{10–52}$$

and

$$0.9 = 1 - \exp\left(\frac{-t_2\omega_T}{h_{FE}}\right) \tag{10–53}$$

giving

$$t_1 = \frac{h_{FE}}{\omega_T}\log_e 0.9 = 0.1\,\frac{h_{FE}}{\omega_T} \tag{10–54}$$

and

$$t_2 = \frac{h_{FE}}{\omega_T}\log_e 0.1 = 2.3\,\frac{h_{FE}}{\omega_T} \tag{10–55}$$

Hence,

$$t_r = t_2 - t_1 = 2.2\,\frac{h_{FE}}{\omega_T} \tag{10–56}$$

It is seen from Eq. (10–56) that, for a given large-signal gain, a high value of f_T produces a low rise time.

There are three major omissions in this very simple analysis. It does not account for variations in ω_T and h_{FE} with voltage and current changes; it does not consider the effects of a load in the collector circuit; it is not directly applicable when the transistor goes into saturation. The waveforms which are obtained experimentally when switching the transistor on and off when saturation is reached are given in the next section: a more complete analysis which takes the omissions into account is presented in Sec. 10–10.

10–9 Definition of Switching Times for the Transistor: Saturation Case

A common method of pulsing the transistor is to control the base current in a CE circuit so that the transistor can be turned on or off. Figure 10–15(a) shows a schematic circuit that will produce a suitable base current waveform. When the switch is in position A, a constant current, I_{b_1}, flows in the base lead, and when the switch is in position B the base-emitter junction is reverse biased. If I_{b_1} is high enough to saturate the transistor, the collector current waveform of Fig. 10–14(b) is obtained. There are four time intervals indicated on the figure:

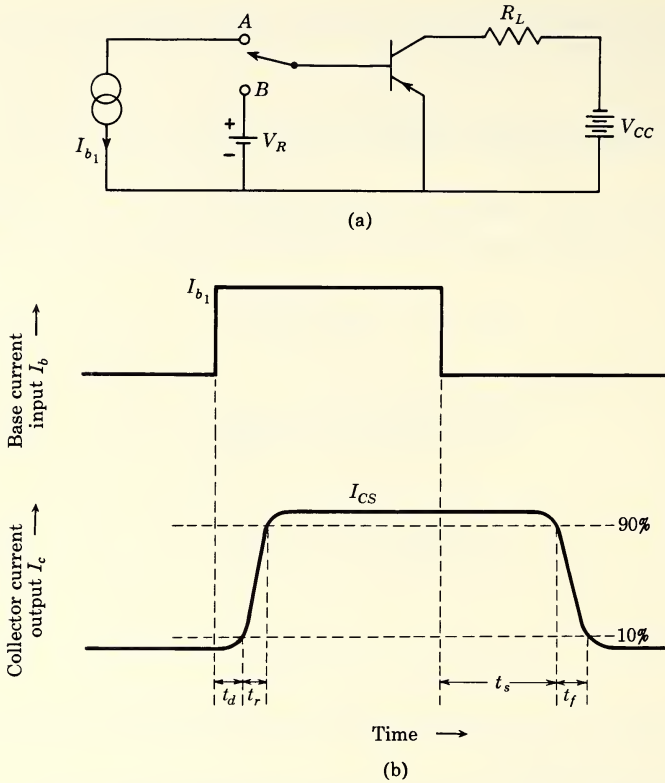


FIG. 10–15. (a) Switching circuit for a *pnp* transistor. (b) Input and output current waveforms when saturation achieved.

t_d, the *delay time*, is the time which elapses between the onset of base current and the rise of the collector current to 10 per cent of its final value.

t_r, the *rise time*, is the time it takes the collector current to increase from 10 per cent to 90 per cent of its final value.

t_s, the *storage time*, is the time required to remove excess charge (Q_{BX} in Fig. 10–11) from the base region before the transistor comes out of saturation. It is defined as the time which it takes the collector current to fall to 90 per cent of its saturation value after the base current is switched off.

t_f, the *fall time*, is the time for the collector current to drop from 90 per cent to 10 per cent of its saturation value.

If the transistor is not allowed to saturate, the storage time is very small. Since the total time required to switch the transistor off is $t_s + t_f$, it would appear that the transistor should never be operated in saturation for fast switching off. However, many applications arise where it is desirable to have

the "on" condition of the transistor in saturation since this usually results in more simple circuitry and less power loss in the device. Overdriving the base circuit, which is necessary for saturation, also has the advantage that it reduces the rise and fall times as shown in Sec. 10–11.

10–10 Charge Control of the Transistor

In this section, the elementary concept of *base charge control* of the transistor is developed. This method of analysis leads to a greater understanding of the operation of the transistor and is of great value in predicting switching times.

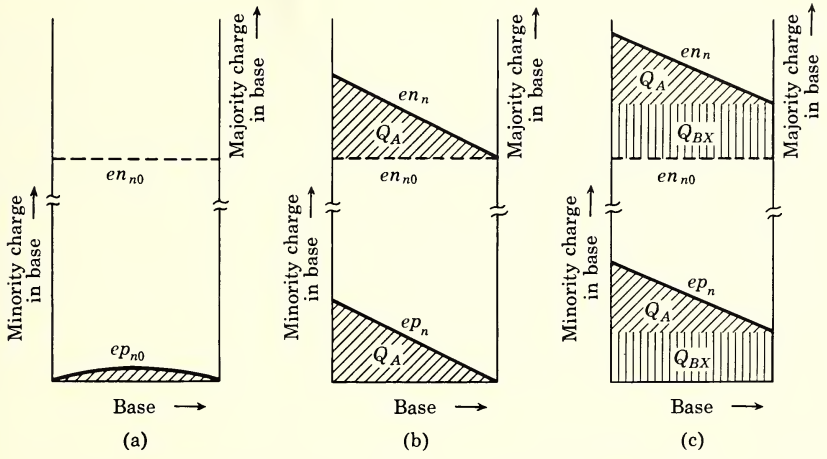


FIG. 10–16. Minority and majority charge in a diffusion *pnp* transistor under three conditions of operation.

Figure 10–16 gives the minority and majority charge distributions in the base region of a *pnp* diffusion transistor under three conditions of operation. When both the collector-base and emitter-base junctions are reverse biased, part (a) of the figure shows that the minority carrier density is small over the whole width of the base and that it falls off close to the two reverse-biased junctions. Under small-current conditions the majority charge density is assumed to be unchanged across the base region as shown in Fig. 10–16(a). Assuming that no acceptor atoms are present in the *n*-type base material, we can write

$$n_{n0} = p_{n0} + N_d \approx N_d \tag{10–57}$$

since $n_{n0} \gg p_{n0}$ in practice. The majority carrier density in the base is constant at approximately the value N_d.

Figure 10–16(b) shows the base charge when the transistor is operating with a high emitter current. This current is injected into the base from

the emitter region and substantially increases the minority charge density close to the emitter-base junction. The current is carried across the base entirely by diffusion, and so the minority charge drops linearly with distance, falling to a value of almost zero at the collector-base junction. Because of charge neutrality, the majority charge density has the profile shown in Fig. 10–16(b), that is, it is everywhere increased above its thermal equilibrium value by an amount equal to the increase in the minority charge density. (See Sec. 9–7 for a brief discussion of charge neutrality under these conditions.) We can write,

$$n_n = p_n + N_d \tag{10–58}$$

and so from Eq. (10–57),

$$n_n - n_{n0} = p_n - p_{n0} \tag{10–59}$$

The final diagram, Fig. 10–16(c), shows the base charge density when the transistor is in saturation. In this case, the minority charge at the collector end of the base region is not zero, and there is a large amount of minority charge stored in the base region. Equation (10–59) is still obeyed, giving the charge profile shown.

When the transistor is switched from the "off" to the "on" state, the charge contained in the base region must be changed from the value shown

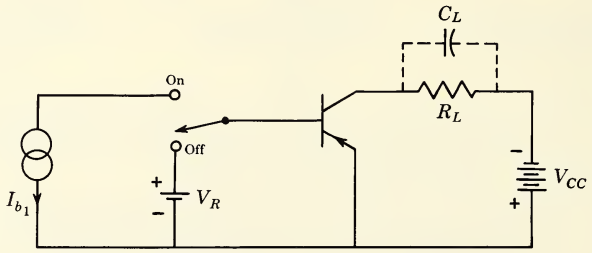


FIG. 10–17. Circuit for switching the transistor on.

in Fig. 10–16(a) to the value shown in part (b). The transient operation of the transistor is studied here by analyzing the way in which the charge in the base region changes from one state to another. In our analysis, we shall neglect the effects of transit time and concentrate on the times required to build up or remove the base charges Q_A and Q_{BX}.

When a *pnp* transistor in the CE connection is switched "on," the current in the base lead results in a flow of *electrons* into the (*n*-type) base region. The *majority* charge density is increased by the current flow in the base lead since no appreciable electron flow occurs at the two junctions. To obtain the collector current, I_c, appropriate to the conditions shown in Fig. 10–16(b), sufficient majority charge (electrons) must have entered the base region (via the base lead) to build up the required charge profile. This is the dominant

factor in determining the rise time. As the *majority* charge in the base region increases, the *minority* charge rises so that charge neutrality is still obeyed. The increase in minority charge density results in the flow of emitter current (i.e., holes across the emitter-to-base junction) and so there is little delay between the rise in majority charge density and the corresponding rise in the minority charge density. The collector current rises in unison with the increase in minority and majority charge densities until the final collector current is reached.

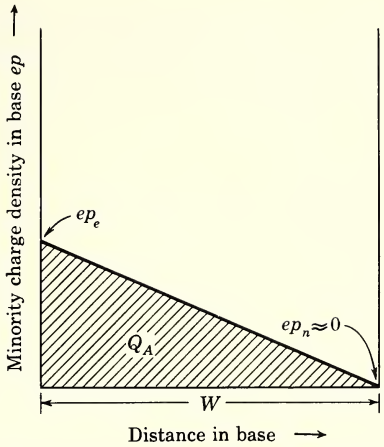


FIG. 10–18. Active charge in the base region for a diffusion transistor.

The rise time for the collector current can be determined by computing the amount of charge which is supplied to increase the active majority charge, Q_A. All of the charge supplied by the base current is not used for this purpose, however, because the junction capacitances and the recombination mechanism require additional charge. For the circuit of Fig. 10–17, the base current supplies five components of charge during the rise time of the collector current. These five components are listed and discussed below.

1. Active charge in the base region, Q_A.
2. Charge on the emitter-base junction capacitance necessary to change the junction potential to its operating value.
3. Charge on collector-base capacitance to correspond with the change in V_{cb}.
4. Charge to replenish that part of the active charge lost by recombination.
5. Charge required to change the voltage across the stray capacitance of the load (see Fig. 10–17).

1. The active charge in the base of a diffusion transistor is shown in Fig. 10–18. Neglecting recombination, the emitter current for a base of area A is

$$I_e \approx I_c = eAD_p \frac{p_e}{W} \tag{10–60}$$

where p_e is the hole density at the emitter end of the base region. The active charge from Fig. 10–18 is

$$Q_A = \frac{ep_eAW}{2} = \frac{W^2}{2D_p} I_c \tag{10–61}$$

and so the amount of active charge in the base region is proportional to the flow of collector current. Assuming $W^2/2D_p$ is constant (i.e., neglecting base modulation),

$$dQ_A = \frac{W^2}{2D_p} dI_c \tag{10–62}$$

From Eqs. (4–45) and (5–15), we can write

$$C_D r_\epsilon = \frac{W^2}{2D_p} \tag{10–63}$$

Hence,

$$dQ_A = C_D r_\epsilon \, dI_c \tag{10–64}$$

This value will be used in the next paragraph.

2. As soon as the collector current starts to flow, the base-to-emitter voltage has a small magnitude which will increase throughout the rise time. During this period, charge is being supplied to the base-to-emitter capacitance, C_{je}. Neglecting recombination in the base, an increase in emitter current, dI_e, causes a change in emitter-to-base voltage of $r_\epsilon \, dI_e$, and so the increased charge on the capacitance is

$$dQ_{je} = C_{je} r_\epsilon \, dI_e \approx C_{je} r_\epsilon \, dI_c \tag{10–65}$$

The sum of the charges in paragraphs 1 and 2 gives

$$\begin{aligned} dQ_A + dQ_{je} &= (C_D + C_{je}) r_\epsilon \, dI_c \\ &= \frac{1}{\omega_T} dI_c \end{aligned} \tag{10–66}$$

from Eq. (10–41).

3. A change in collector current of dI_c causes a voltage change of $R_L \, dI_c$ at the collector terminal. The change in charge on the collector-to-base junction capacitance, C_{jc}, is therefore

$$dQ_{jc} = C_{jc} R_L \, dI_c \tag{10–67}$$

assuming $V_{be} \ll V_{ce}$.

4. When a transistor is operated with a constant current I_c, the base current is given by I_c/h_{FE} where h_{FE} is the large-signal value of the current gain. Thus a base current of I_c/h_{FE} must be supplied to the base lead to replenish the active charge lost by recombination. In a time dt, the charge supplied to the base for this purpose is dQ_R, where

$$dQ_R = \frac{I_c}{h_{FE}} dt \tag{10–68}$$

5. If the load resistance is shunted by the stray capacitance C_L as shown in Fig. 10–17, a change in collector voltage of $R_L I_c$ requires a change in charge

on the capacitance of $C_L R_L \, dI_c$. This charge is supplied from the collector-to-emitter circuit. Therefore, the charge which has to be supplied to the base region to change the charge on the load capacitance by $C_L R_L \, dI_c$ is

$$dQ_c = \frac{C_L R_L}{h_{FE}} \, dI_c \tag{10–69}$$

All five components of charge are supplied by the base current, therefore

$$I_b \, dt = dQ_A + dQ_{je} + dQ_{jc} + dQ_R + dQ_c \tag{10–70}$$

Substituting from Eq. (10–66) through (10–69),

$$I_b = \frac{I_c}{h_{FE}} + \left(\frac{1}{\omega_T} + C_{jc} R_L + \frac{C_L R_L}{h_{FE}} \right) \frac{dI_c}{dt} \tag{10–71}$$

In this equation, h_{FE}, ω_T, and C_{jc} are functions of I_c. If we assume that at $t = 0$, I_b rises instantaneously to a value I_{b_1}, then over a small range of values I_c such that h_{FE}, ω_T, and C_{jc} can be considered constant, Eq. (10–71) can be solved to give

$$I_c = h_{FE} I_{b_1} \left[1 - \exp\left(\frac{-t}{h_{FE} \tau} \right) \right] \tag{10–72}$$

where

$$\tau = \frac{1}{\omega_T} + C_{jc} R_L + \frac{C_L R_L}{h_{FE}} \tag{10–73}$$

Equation (10–72) can be compared with the solution for the simple hybrid-π circuit given in Eq. (10–50).

The time constant, τ, will vary as the collector current rises. In general, τ is not a strong function of I_c and, as a first order approximation, we may write,

$$\tau_{RE} = \bar{\tau} \tag{10–74}$$

where $\bar{\tau}$ is the mean value of τ. Experimentally, τ_{RE} can be measured for a particular transistor by finding the rise time under given conditions. (See Fig. 10–20 in Sec. 10–11 for a curve of τ_{RE} measured under saturation conditions and for comments on the variation in τ_{RE} with I_{CS}.) From Eqs. (10–50) and (10–56),

$$t_r = 2.2 h_{FE} \tau_{RE} \tag{10–75}$$

where h_{FE} is the mean value of the current gain over the current range concerned.

It is possible to calculate the variation in the individual terms of τ_{RE} as I_c or V_{CC} is varied. This can be done if the curves of ω_T, h_{FE}, and C_{jc} are known. The method can then be used to provide a specification of the transistor under pulsed conditions of operation.

The fall time of the transistor is almost the same as the rise time since the transistor is operating over the same part of its characteristics. The delay

time is composed of two time periods. The first of these is the time elapsing between the switching on of the base current and when the base-to-emitter voltage just reaches the forward direction. This time was calculated in Sec. 10–8. To this must be added the time taken for the collector current to rise to 10 per cent of its final value, which can be calculated from a knowledge of τ_{RE}.

From this brief account, it can be seen that this is far from being an exact method of analysis. Transistors have so many interrelated parameters, however, that approximations must always be made and errors of up to 20 per cent are permissible. Further details of the transient response characterization of transistors can be found in the two papers by Ekiss and Simmons cited in the bibliography at the end of the chapter.

10–11 Switching Times for a Saturated Transistor

The transistor shown in Fig. 10–17 will saturate if

$$h_{FE}I_{b_1} > I_{CS} \approx \frac{-V_{CC}}{R_L} \tag{10–76}$$

In this case, the rise time of the collector current is the time elapsing between the attainment of 10 per cent and 90 per cent of the saturation current I_{CS}. From Eq. (10–72), and substituting τ_{RE} for the average time constant, we have

$$0.1I_{CS} = h_{FE}I_{b_1}\left[1 - \exp\left(\frac{-t_1}{h_{FE}\tau_{RE}}\right)\right] \tag{10–77}$$

and

$$0.9I_{CS} = h_{FE}I_{b_1}\left[1 - \exp\left(\frac{-t_2}{h_{FE}\tau_{RE}}\right)\right] \tag{10–78}$$

Thus,

$$t_r = t_2 - t_1 = h_{FE}\tau_{RE}\log_e\left[\frac{h_{FE}I_{b_1} - 0.1I_{CS}}{h_{FE}I_{b_1} - 0.9I_{CS}}\right] \tag{10–79}$$

For the case when the transistor is driven well into saturation so that $h_{FE}I_{b_1} > 5I_{CS}$, Eq. (10–79) can be written in an approximate form by using a series expansion and retaining only the significant terms. Then,

$$t_r = 0.8\tau_{RE}\frac{I_{CS}}{I_{b_1}} \tag{10–80}$$

Figure 10–19 shows the rise in collector current when the transistor is well in saturation. The collector current follows the dotted line of the initial slope until I_{CS} is reached. Increasing the base drive (I_{b_1}) for a given saturation current decreases the rise time.

The variation of τ_{RE} with I_{CS} is shown in Fig. 10–20 for a medium frequency, alloy-junction transistor. The curve was obtained by measuring the

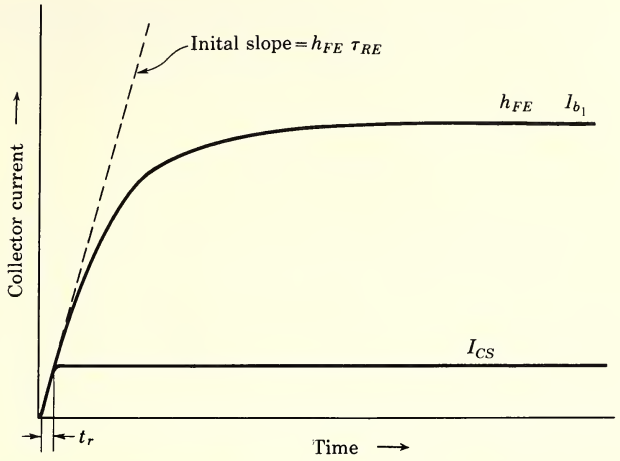


FIG. 10-19. Rise of collector current when transistor is well in saturation. Rise time, t_r, is time between 10 per cent and 90 per cent I_{CS}.

rise time for the saturated case as a function of I_{CS}. The ratio I_{CS}/I_{b_1} was constant throughout the measurements. At small values of I_{CS}, τ_{RE} is high because the term $C_{jc}R_L$ dominates Eq. (10-73). As I_{CS} increases, τ_{RE} diminishes because ω_T increases and the other terms decrease. When I_{CS} is large, τ_{RE} slightly increases again. At the vicinity of the broad minimum of τ_{RE}, the term $C_{jc}R_L$ is typically about 10 per cent of the total value. R_LC_L/h_{FE} is assumed to be negligible in this case.

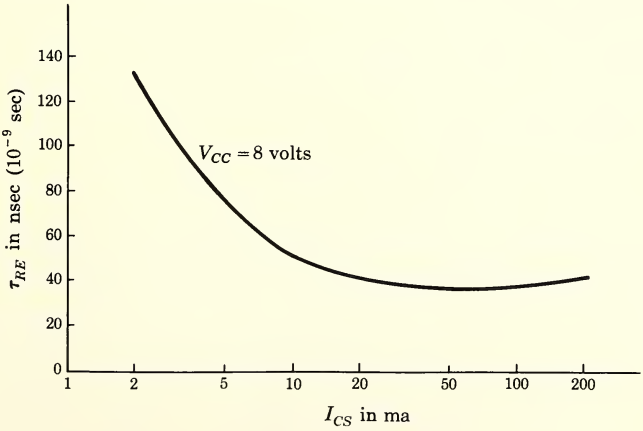


FIG. 10-20. Variation in τ_{RE} with I_{CS} for a low frequency transistor. [From J.A. Ekiss and C.D. Simmons, "Junction Transistor Transient Response Characterization," *Solid State Journal*, **2**, 1 (January 1961), p. 20.]

When a saturated transistor is switched off, the excess charge Q_{BX} (shown in Fig. 10–11) must be removed from the base region before the collector current will fall again. The base current which will just saturate the transistor is

$$I_{BS} = \frac{I_{CS}}{h_{FE}} \tag{10–81}$$

Therefore, the transistor may be brought out of saturation (though it will not be switched "off") if the base current is dropped below I_{BS}. If the base current is dropped to zero, the transistor will eventually return to the "off" state. However, there is a third possibility. This is that the base current may be reversed for a short time to remove the charge in the base region more quickly and so reduce the storage time. All three possibilities are accounted for in the following equations.

Let Q be the total charge in the base region at any time, t, where Q includes Q_A and Q_{BX} of Fig. 10–11. When a current I_b is flowing in the base lead, we may write

$$\frac{dQ}{dt} = I_b - \frac{Q}{\tau_s} \tag{10–82}$$

where τ_s is the *lifetime* of the charge carriers in the base region, and we have neglected the charges on the capacitances. When a current I_{b_1} is flowing and the transistor is saturated, $dQ/dt = 0$, and

$$I_{b_1}\tau_s = Q = Q_A + Q_{BX} \tag{10–83}$$

We shall now examine the conditions during the storage time. Let base current I_{b_2} flow *in the same direction as* I_{b_1}. Equation (10–82) can now be written as

$$\frac{dQ}{dt} + \frac{Q}{\tau_s} = I_{b_2} \tag{10–84}$$

and a solution is of the form

$$Q = I_{b_2}\tau_s + C \exp\left(\frac{-t}{\tau_s}\right) \tag{10–85}$$

where C is a constant to be determined. Assume that I_{b_2} starts to flow at $t = 0$ and that up to this time current I_{b_1} was flowing. Then by comparing Eqs. (10–83) and (10–85), at $t = 0$,

$$I_{b_1}\tau_s = Q = I_{b_2}\tau_s + C$$

giving,

$$C = (I_{b_1} - I_{b_2})\tau_s \tag{10–86}$$

Equation (10–85) now becomes

$$Q = I_{b_2}\tau_s + (I_{b_1} - I_{b_2})\tau_s \exp\left(\frac{-t}{\tau_s}\right) \tag{10–87}$$

The transistor just comes out of saturation when $Q_{BX} = 0$ and $Q = Q_A$. From Eqs. (10–81) and (10–82),

$$Q_A = I_{BS}\tau_s \tag{10–88}$$

Hence,

$$I_{BS}\tau_s = I_{b_2}\tau_s + (I_{b_1} - I_{b_2})\tau_s \exp\left(\frac{-t_s}{\tau_s}\right)$$

or

$$t_s = \tau_s \log_e \left(\frac{I_{b_1} - I_{b_2}}{I_{BS} - I_{b_2}}\right) \tag{10–89}$$

where t_s is the *storage time.*

Equation (10–89) is only approximately obeyed in practice. The lifetime, τ_s, is the value applying to the saturated condition of operation of the transistor. It is usually measured by finding the storage time, t_s, when

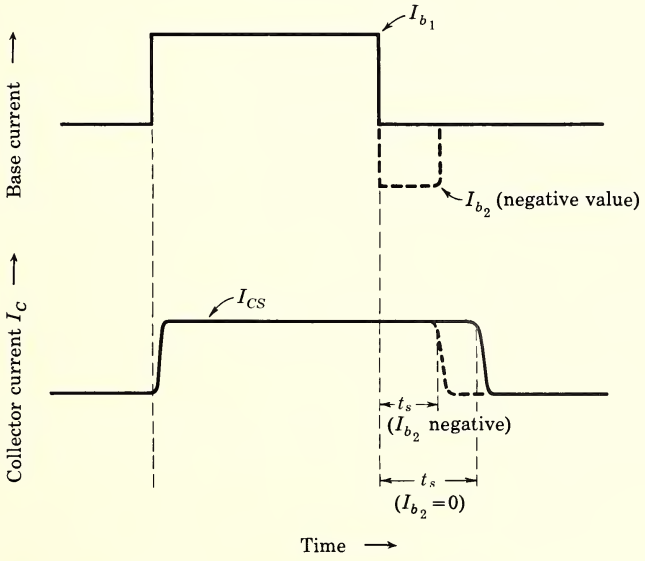


FIG. 10–21. Storage time when $I_{b2} = 0$ and when I_{b2} is negative.

I_{b_1}, I_{b_2}, and I_{BS} are known. τ_s is usually below a microsecond for high frequency transistors and varies somewhat with I_{BS}.

The effect of changing the switching-off base current, I_{b_2}, is illustrated in Fig. 10–21. The full line shows the condition $I_{b_2} = 0$, and the dotted line shows that the storage time is reduced when I_{b_2} has a negative value as can be seen from Eq. (10–89).

The circuit of Fig. 10–22 illustrates the use of a "speed-up" capacitor C. When an input step voltage is applied, the initial base current is high, and the rise time of the circuit is reduced. This type of circuit can also be used to

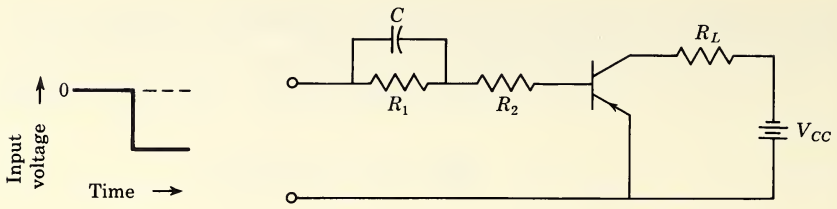


FIG. 10–22. Use of "speed-up" capacitor to reduce rise time.

reduce storage time by effectively removing the stored base charge when the transistor is switched off.

BIBLIOGRAPHY

Hybrid–π circuit

Cote, Alfred J., Jr. and J. Barry Oakes, *Linear Vacuum-tube and Transistor Circuits*, New York: McGraw-Hill Book Company, Inc., 1961

Joyce, Maurice V., and Kenneth K. Clarke, *Transistor Circuit Analysis*, Reading, Mass.: Addison-Wesley Publishing Company, Inc., 1961

Pettit, Joseph Mayo, and Malcolm Myers McWhorter, *Electronic Amplifier Circuits, Theory and Design*, New York: McGraw-Hill Book Company, Inc., 1961

Pulse operation

DeWitt, David and Arthur L. Rossoff, *Transistor Electronics*, New York: McGraw-Hill Book Company, Inc., 1957

Ekiss, J. A., and C. D. Simmons, "Junction Transistor Transient Response Characterization," Parts I and II, *Solid State Journal*, **2**, 1 (January 1961), 17–24; **2**, 2 (February 1961), 24–29

PROBLEMS

10–1 Calculate the Miller capacitance C for the transistor of Prob. 9–4. What is the gain-bandwidth product for this transistor?

10–2 Find the collector load resistance for the hybrid-π circuit of Fig. 10–1 at which $\omega_{3db} = \omega_\beta/2$.

ans. 3900 ohms

10–3 Fig. 10–1 shows the value of the components of the hybrid-π circuit when $I_E = 1$ ma and $V_{CE} = -5$ volts. What are the component values when $I_E = 10$ ma and $V_{CE} = -10$ volts? What effect does this have on ω_o?

10–4 A manufacturer makes a diffused base transistor which can be adequately represented by a modified hybrid-π circuit where

$$r_{b'e} = \frac{r_\epsilon}{1 - \alpha_0 \exp(-j0.5\omega/\omega_\alpha)}$$

$$g_m = \frac{\alpha_0}{r_\epsilon} \exp\left(\frac{-j0.5\omega}{\omega_\alpha}\right)$$

Assuming that the frequency of operation is considerably smaller than ω_T, show that there is a frequency range where the gain of the transistor drops at the rate of 6 db per octave and compute the phase shift through the device as a function of frequency.

10–5 How would you expect the value of ω_T to vary with V_{CE} for an alloy junction transistor? Give a physical reason for all the factors which are discussed and briefly define all new terms that are introduced.

10–6 A CE transistor stage can be represented by a hybrid-π circuit. Assign typical values to all the elements and then consider how this equivalent circuit can be used to determine gain and frequency response of a multistage amplifier. Calculate these two quantities using typical values of resistances and capacitances. (a-c circuit only required.)

10–7 Explain how the methods of charge control can be applied to the understanding and specification of transistors for pulse operation.

10–8 Using the information given in Secs. 10–10 and 10–11, calculate the rise time for a transistor having the τ_{RE} characteristic shown in Fig. 10–20, under the following conditions: (a) $I_{b_1} = 10$ ma, $V_{CC} = -8$ volts, $R_L = 320$ ohms, (b) $I_{b_1} = 10$ ma, $V_{CC} = -8$ volts, $R_L = 4000$ ohms. If h_{FE} is considered to be a constant of value 50, what is the maximum capacitance that can be added across the load to increase the rise time by 10 per cent in both cases?

10–9 The transistor of Prob. 10–7 is operating in saturation with $V_{CC} = -8$ volts, $R_L = 320$ ohms and $I_{b_1} = 10$ ma. When the base current is suddenly reduced to zero, the storage time is 200 nsecs. What is the value and direction of the base current necessary to reduce this time to 100 nsecs?

11

The Tunnel Diode

The tunnel diode is an important solid-state device that was discovered by Esaki in 1958. It consists of a heavily doped p–n junction in which the transition from p- to n-type material occurs over a distance of less than 10^{-6} cms (100 A). As is shown later in this chapter, this device can be used as an amplifier, an oscillator or a switch. The tunnel diode has many applications in fast electronic circuitry since it is capable of operating at a frequency in excess of 10^{10} cycles per second and of switching in a time less than 10^{-9} seconds. Among the advantages of

the tunnel diode are its small size, its low voltage and power requirements, a very wide temperature range of operation, low cost and ease of fabrication, and its extremely high speed of operation. The principal disadvantage of the device is that it has only two terminals, and so its equivalent circuit over part of its characteristic is a one port active network. Thus, the input and output circuits cannot be isolated and precautions must be taken to avoid unwanted high frequency oscillations.

The tunnel diode gets its name from the "tunnel" effect which is a phenomenon that can only be explained by quantum mechanics. Accordingly, the first section of this chapter gives a brief quantum mechanical picture of an electron "tunneling" through a potential barrier.

11-1 Quantum Mechanical Tunneling

In classical physics, a particle with total energy E cannot pass into a region where the potential energy, U, has a higher value. Thus, a ball of mass m with kinetic energy $\frac{1}{2}mv^2$ can travel up a hill to a height h given by $\frac{1}{2}mv^2 = mgh$,

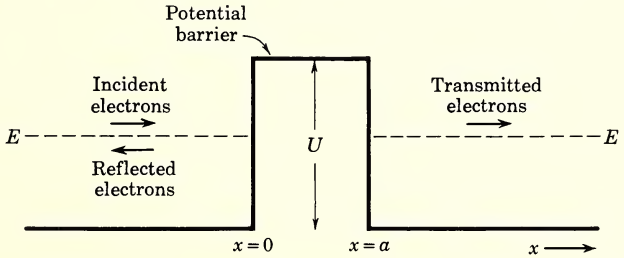


FIG. 11-1. Diagram of electrons incident on a potential barrier.

where g is the acceleration due to gravity, but cannot exceed that height. However, many phenomena in nuclear and atomic physics can only be explained by assuming penetration of such a high barrier by a subatomic particle. Using quantum mechanics, it can be shown that there is a small but finite probability of a particle passing through a higher energy barrier. This is known as the *tunnel effect*.

Figure 11-1 shows a beam of electrons of energy E incident on a barrier of potential energy U. As shown in the diagram, $E < U$. By referring back to Sec. 1-8, we see that the behavior of the incident electrons can be described by the wave function, ψ. The wave function in the x direction is the solution of the Schrödinger wave equation [Eq. (1-15)]. This equation is repeated here for convenience.

$$\frac{d^2\psi}{dx^2} + \frac{8\pi^2 m}{h^2}(E - U)\psi = 0 \qquad (11\text{-}1)$$

Particles traveling in the positive x direction in free space may be represented by waves having the alternative forms:

$$\psi = A \exp [j(\omega t - cx)] \tag{11–2}$$

or

$$\psi = A \exp [-j(\omega t - cx)] \tag{11–3}$$

We will use the second form and suppress the time-dependent term. Electrons traveling in the positive x direction will, therefore, be represented by

$$\psi = A \exp (jcx) \tag{11–4}$$

Electrons moving in the negative x direction can be represented by

$$\psi = B \exp (-jcx) \tag{11–5}$$

On the left-hand side of the potential barrier of Fig. 11–1, we can postulate that A_1 is the amplitude of the incident wave and that B_1 is the amplitude of

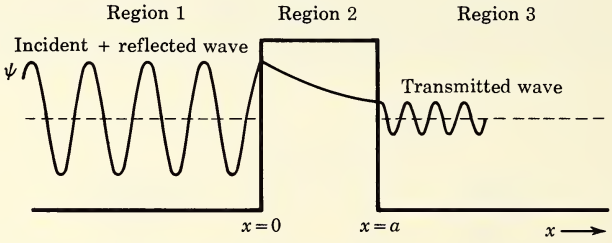


FIG. 11–2. Diagrammatical representation of tunneling.

the wave reflected from the potential barrier. We can, therefore, write the wave function in the form

$$\psi_1 = A_1 \exp (jc_1x) + B_1 \exp (-jc_1x) \tag{11–6}$$

where

$$c_1^2 = \frac{8\pi^2 mE}{h^2} \tag{11–7}$$

since $U = 0$ in this region. The energy of the incident electrons (E) is positive, and so c_1 is real and the wave motion is as represented in region 1 ($x < 0$) in Fig. 11–2.

The potential barrier exists between $x = 0$ and $x = a$ and is designated as region 2. Here the value of $E - U$ is negative, and so the exponents in the solution of the wave equation are real. Writing

$$c_2^2 = \frac{8\pi^2 m}{h^2} (U - E) \tag{11–8}$$

where c_2 is defined in such a way as to be real, we have

$$\psi_2 = A_2 \exp (c_2x) + B_2 \exp (-c_2x) \tag{11–9}$$

We note that in many practical applications (such as the tunnel diode), $c_2 a \gg 1$. Hence, $A_2 \ll B_2$ since $\psi \to 0$ as $a \to \infty$. Note that the electrons cannot be detected in this region since the wave function is not oscillatory.

In region 3, $x > a$, $U = 0$ and there will be a transmitted wave only in the positive direction of x. Therefore, the solution of the Schrödinger wave equation is

$$\psi_3 = A_3 \exp(jc_1 x) \tag{11-10}$$

where $c_1 = 8\pi^2 mE/h$ as before.

From Eq. (11-10), we see that A_3 is the amplitude of the wave function representing the transmitted beam of electrons. Provided $A_3 \neq 0$, there will be some probability of electrons "tunneling" through the potential barrier since $|A_3|^2$ is proportional to the number of transmitted electrons (see Sec. 1-8).

The relationship of the amplitudes A_1, B_1, A_2, B_2, and A_3 can be found by applying the conditions that ψ and $d\psi/dx$ are continuous at the boundaries between the regions. This is a necessary mathematical condition relating to Eq. (11-1) since ψ, E, and U are finite. ψ must be finite since it is related to the probability of finding particles at a particular location; E and U are finite in all practical cases. Applying these conditions at $x = 0$ and $x = a$ and eliminating B_1, A_2, and B_2 from the resulting equations allows us to determine the ratio of A_3 to A_1. We will not carry out this procedure here, but it can be shown that the ratio of the intensities of the transmitted and reflected waves, when $A_2 \ll B_2$, gives a transmission coefficient T, where

$$T = \frac{|A_3|^2}{|A_1|^2} = \frac{4 \exp(-2c_2 a)}{1 + \frac{1}{4}(c_2/c_1 - c_1/c_2)^2} \tag{11-11}$$

Thus, T is the fraction of the electrons in the incident beam "tunneling" through the barrier.

Equation (11-11) shows that there is a finite but small probability that particles will cross a potential barrier even though their initial energy is smaller than the barrier height. This probability decreases exponentially with the width of the barrier (a) and also decreases as the barrier height increases [Eq. (11-8)]. Figure 11-2 shows the wave functions diagrammatically in the three regions. It should be noted that both the incident and the transmitted electrons have the same energy E. Tunneling will only take place when electrons possess the same energy on each side of the barrier.

11-2 Energy Bands of the Tunnel Diode

The energy bands of a tunnel diode can be obtained by reference to those of a conventional p-n junction. If the acceptor and donor densities on the two sides of a junction are less than about 10^{18} cm^{-3}, Fig. 11-3(a) applies. On

each side, the Fermi level (defined as that energy where the probability of occupation is one half) lies in the forbidden gap (see Secs. 2–4 and 2–5).

Figure 11–3(b) shows the energy bands for the tunnel diode. On the *n*-type side, the donor density is greater than 10^{19} cm^{-3}, and thus a large number of electrons is present in the conduction band. The Fermi level for this case lies within the conduction band. Similarly, the acceptor atoms on the *p*-type side are so numerous that there are many holes in the valence band and the Fermi level is within that band.

At room temperature, the presence of the Fermi level within the conduction band of the *n*-type material means that most of the electron states at the bottom of the band will be occupied. The distribution of filled states can be calculated from Fermi-Dirac statistics (see Chap. 2). However, the Boltzmann factor that was used in Chap. 2 as an approximation to the

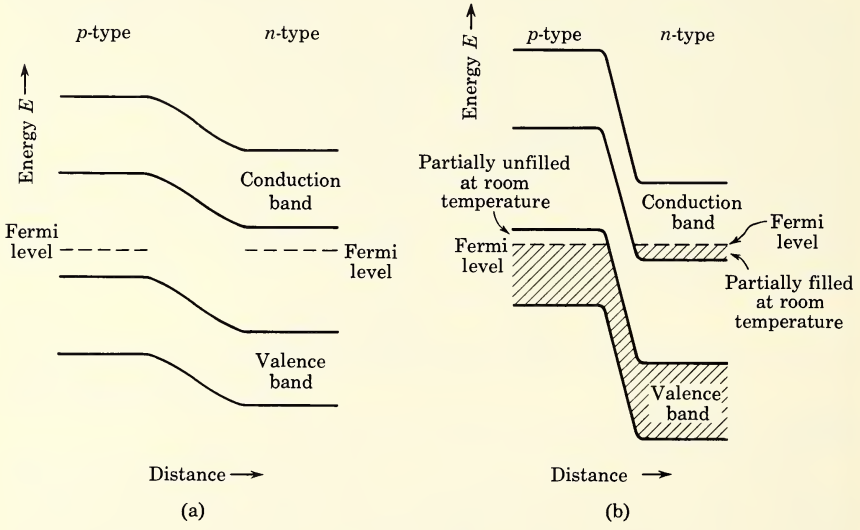


FIG. 11–3. (a) Energy bands for a conventional *p-n* junction (zero applied voltage and room temperature). (b) Energy bands for the tunnel diode where both impurity densities are greater than 10^{19} cm^{-3} (zero applied voltage and room temperature).

Fermi-Dirac distribution is no longer applicable since there are now states close to the Fermi level. Thus, the Fermi factor as given in Eq. (2–10) must be used. This is a condition known as *degeneracy*. By this definition, it is seen that both the *n*- and *p*-type material used in a tunnel diode are degenerate.

Whatever the temperature, the Fermi levels on the two sides of the junction will line up. For the tunnel diode, therefore, tunneling will take place where the two bands overlap in their energy ranges. This has been indicated on Fig. 11–4. Here, it is possible for electrons to tunnel through the potential

barrier separating the conduction band of the *n*-type material from the valence band of the *p*-type material. The tunneling current will be proportional to the product of the probability of occupancy on the side the electrons leave, the probability of nonoccupancy on the side they reappear, and the transition probability. When there is no applied voltage, the net tunneling current is zero since the probability of occupation on the two sides is the same at the same energy level and the probability of tunneling in either direction is the same.

A calculation of the tunnel current flowing in either direction can be made by the application of quantum mechanics. The analysis must take into account the variation in the height of the potential barrier with distance, since

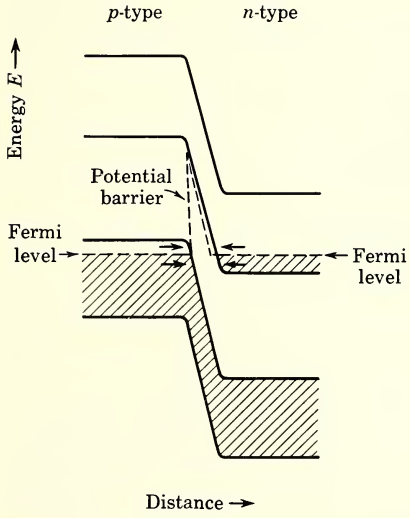


FIG. 11-4. Energy bands of the tunnel diode showing tunneling at room temperature and zero applied voltage.

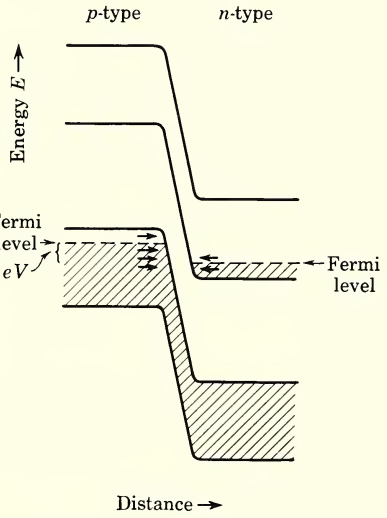


FIG. 11-5. Energy bands of the tunnel diode at room temperature and reverse bias condition.

the barrier is formed by the lower edge of the conduction band across the junction (shown dotted). For tunneling to be appreciable, the width of the junction must be small. This condition can be achieved in an abrupt *p-n* junction which is doped to degeneracy on both sides of the junction.

11-3 Forward and Reverse Characteristics

When a voltage, V, is applied to the junction, the Fermi levels on the two sides are spaced eV apart on the energy band diagram. Figure 11-5 shows the case where V is negative. Here, the reverse voltage has depressed the Fermi level on the *n*-type side by an energy eV relative to that of the *p*-type Fermi level.

The number of electrons which can tunnel to the right is much increased from the case where $V = 0$. This is because electrons from the numerous occupied states close to the Fermi level on the left-hand side, and above the Fermi level of the right-hand side, can tunnel into the relatively unoccupied states at the same level on the right. Thus, the reverse current of the tunnel diode is high, as shown below in Fig. 11–6.

FIG. 11–6. Detail of energy bands and tunnel diode characteristic.

When the junction is biased in the forward direction with a voltage of a few millivolts, the tunneling current of electrons from *n*- to *p*-type is greatly increased. This is shown in Fig. 11–6(a). In this case, there are many filled states at the bottom of the conduction band of the *n*-type material in line with unfilled states in the valence band of the *p*-type. The number of electrons tunneling in the opposite direction is much smaller.

In Fig. 11–6(b), a higher forward voltage has been applied. Consequently, the amount by which the two bands overlap in energy has been reduced.

There is a corresponding reduction in the tunnel current since some of the filled energy states just below the Fermi level in the *n*-type conduction band are now above the top of the *p*-type valence band. Electrons in these states cannot tunnel since there are no available states on the opposite side to receive them. Therefore, a *further increase* in forward voltage results in a *decrease* in tunneling current in the forward direction. As the voltage is increased, tunneling ceases and eventually the conventional injection current of the *p-n* junction predominates as shown in Fig. 11–6(c).

The characteristic curve given in Fig. 11–6(d) shows the values I_p and V_p for the peak tunneling current and voltage, and also I_v and V_v for the current and voltage at the "valley" of the curve. The voltage swing, V_s, is also defined on this characteristic. It is the voltage between the two points on the curve where the current is equal to I_p. The value of V_s is a characteristic of the material as shown in Table 11–1.

TABLE 11–1 Properties of tunnel diodes.

Material	*Peak-to-valley current ratio* I_p/I_v	V_s *(volts)*	*Energy gap* E_g *in ev*
Ge	10–30	0.45	0.72
Si	3–5	0.7	1.1
GaAs	30–60	1.0–1.2	~1.4

The reduction in tunnel current as the voltage is increased results in a *negative resistance* region. In this region, power is transferred from the supply to the external circuit. It is this property of the tunnel diode which makes it useful as an amplifier, an oscillator or a switch (see Sec. 11–4).

Tunnel diodes can be made of the semiconducting elements, germanium or silicon. However, gallium arsenide (GaAs) and other group 3–5 compounds are also used. These are materials where the group 3 and 5 atoms form crystalline compounds having properties similar to the group 4 semiconductors. All types of tunnel diode exhibit basically similar characteristics; for example, the value of the peak current increases with the junction area. Peak currents above one hundred amperes have been obtained in germanium units of large area. The voltage range of the tunnel diode depends on the particular semiconductor used, and little variation is possible for any one type of material. Values of the peak-to-valley current ratio, the voltage swing, and the energy band gap for three types of tunnel diode are tabulated in Table 11–1. It is seen that the negative resistance region occurs below one volt for all three types.

Since tunnel diodes are fabricated with very heavily doped semiconductors, it is not absolutely necessary to use very pure material in their manufacture. In addition, in many cases, surface conditions are much less critical than in

transistor fabrication. Thus, many types of tunnel diodes are easier and cheaper to manufacture than transistors. However, special types of tunnel diodes with tight parameter tolerances and special performance properties can be as expensive as any other semiconductor device.

The negative resistance characteristic of the tunnel diode may be observed over quite a wide temperature range. Both high and low temperature operation are possible since the device is heavily doped on both sides of the junction and no appreciable (temperature-sensitive) minority carrier injection occurs during tunneling. For germanium units, a temperature range of 4.2° K to 250° C has been observed; for silicon, the upper temperature limit is extended to about 350° C. The peak-to-valley current ratio is temperature sensitive and varies widely among different units.

11–4 The Tunnel Diode as a Circuit Element

There are two important factors which govern the use of the tunnel diode in electronic circuits. In the first place, the limitations in switching time and frequency of operation are those imposed by the equivalent and external

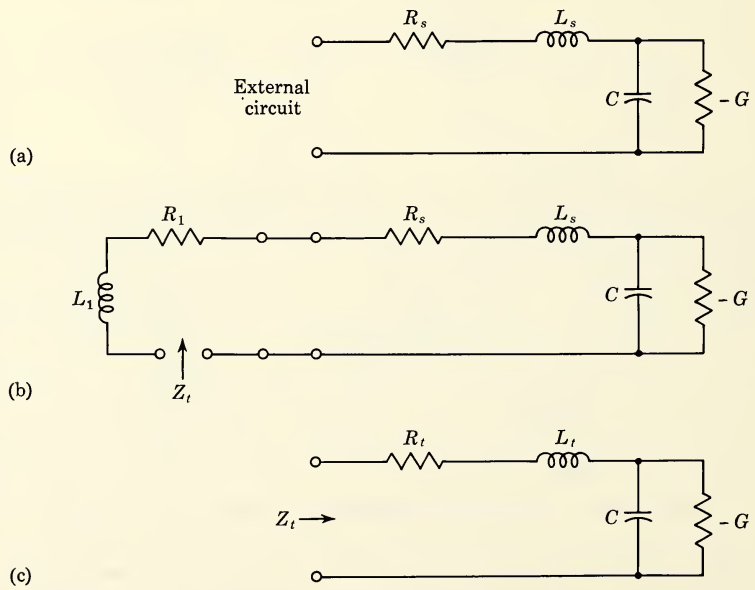


FIG. 11–7. (a) Equivalent circuit of tunnel diode. (b) Tunnel diode and external circuit. (c) Equivalent circuit of the combination.

circuit of the diode, and the inherent limit set by the tunneling mechanism is never approached. Nevertheless, switching times of less than 10^{-10} seconds, and oscillation frequencies above 10^{10} cycles per second are possible in units

having high peak current to capacity ratios, I_p/C (values as high as 200 ma/pf have been achieved in germanium. Secondly, the tunnel diode is a one port (two terminal) network and this raises difficulties in operation since the input and output circuits cannot be separated.

When the tunnel diode is operated in the negative resistance part of its characteristic, the equivalent circuit for small-signal operation is as shown in Fig. 11-7(a). R_s is the ohmic resistance of the semiconducting material, and the construction of the device determines the small series inductance, L_s. C is the junction capacitance due to the depleted charge regions on the two sides of the *p-n* junction. The negative resistance has been represented by a conductance $-G$. Typical values for these quantities are: $R_s = 2$ ohms, $L_s = 10^{-2}$ microhenries, $C = 10$ picofarads, $-G = -0.01$ mhos.

Assume that the tunnel diode is operating with an external circuit having a series resistance R_1 and a series inductance L_1 as given in Fig. 11-7(b). In part (c) of this figure the equivalent circuit of the combination is shown, where $L_t = L_1 + L_s$ and $R_t = R_1 + R_s$. Note that this final circuit is similar to the equivalent circuit of the tunnel diode alone, where R_s is replaced by R_t and L_s is replaced by L_t. Thus, the analysis of the complete circuit applies equally to the tunnel diode under short-circuit a-c conditions when R_s and L_s are substituted for R_t and L_t.

The impedance, Z_t, at an angular frequency ω, is given by

$$Z_t = R_t + j\omega L_t + \frac{1}{(-G) + j\omega C} \tag{11-12}$$

or

$$Z_t = R_t + \frac{(-G)}{G^2 + \omega^2 C^2} + j\left(\omega L_t - \frac{\omega C}{G^2 + \omega^2 C^2}\right) \tag{11-13}$$

Note that both the real (resistive) and imaginary (reactive) parts of the expression are frequency dependent.

There are two important frequencies for this circuit. ω_R is the *resistance cut-off angular frequency* at which the resistive part of Z_t becomes zero. Above this frequency, the tunnel diode cannot amplify. ω_X is the *self-resonant angular frequency* where the imaginary part of Z_t becomes zero. From Eq. (11-13),

$$\omega_R = \left(\frac{G}{C^2 R_t} - \frac{G^2}{C^2}\right)^{1/2} = \frac{G}{C}\left(\frac{1}{R_t G} - 1\right)^{1/2} \tag{11-14}$$

$$\omega_X = \left(\frac{1}{L_t C} - \frac{G^2}{C^2}\right)^{1/2} = \frac{G}{C}\left(\frac{C}{G^2 L_t} - 1\right)^{1/2} \tag{11-15}$$

The system stability may be investigated by determining the distribution of the poles and zeros in the complex S-plane. Writing Z_t in terms of the complex frequency, S, we have

$$Z_t(S) = \frac{S^2 L_t C + (R_t C - L_t G)S + 1 - R_t G}{SC - G} \tag{11-16}$$

The zeros are given by

$$S = -\frac{1}{2}\left(\frac{R_t}{L_t} - \frac{G}{C}\right) \pm \left[\frac{1}{4}\left(\frac{R_t}{L_t} - \frac{G}{C}\right)^2 - \frac{1 - R_t G}{L_t C}\right]^{1/2} \tag{11-17}$$

The condition for stability is that the zeros are in the left-half side of the S-plane. From Eq. (11–17), S will have a negative real part only if

$$\frac{R_t}{L_t} > \frac{G}{C} \tag{11-18}$$

and

$$R_t < \frac{1}{G} \tag{11-19}$$

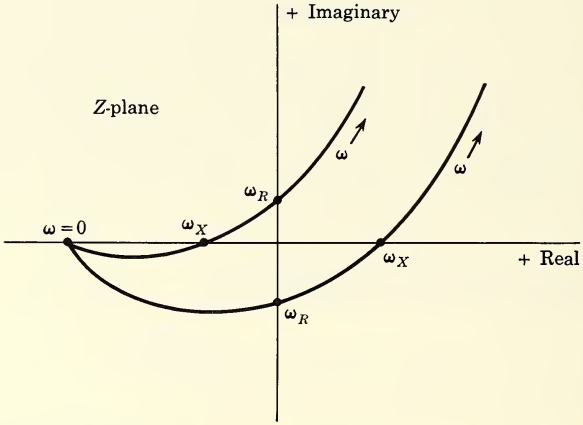


FIG. 11–8. Nyquist diagram for the tunnel diode circuit.

Thus the conditions for stability may be written in the form

$$\frac{1}{G} > R_t > \frac{L_t G}{C} \tag{11-20}$$

The stability conditions when substituted in Eqs. (11–14) and (11–15) give

$$0 < \omega_R < \omega_X \tag{11-21}$$

The Nyquist plot of this case is shown in the lower curve of Fig. 11–8. Physically, we can say that the negative conductance is an energy source which becomes zero at the resistance cut-off angular frequency, ω_R. For stability, this angular frequency must be real and positive ($\omega_R > 0$) and less than the angular frequency where the reactance of the circuit becomes zero ($\omega_R < \omega_X$).

Equations (11–18) and (11–19) give the two requirements for tunnel diode amplifier stability. In terms of the total circuit values, R_t and L_t, the expressions can be rewritten to give

$$R_t < \frac{1}{G} \tag{11-22}$$

and $$L_t < \frac{R_t C}{G} \tag{11-23}$$

It is of interest to see what will happen if these requirements are not fulfilled. For instance, if $R_t > 1/G$, ω_R is no longer real, and the device cannot be used as an amplifier but may be used as a switch. When $L_t >$

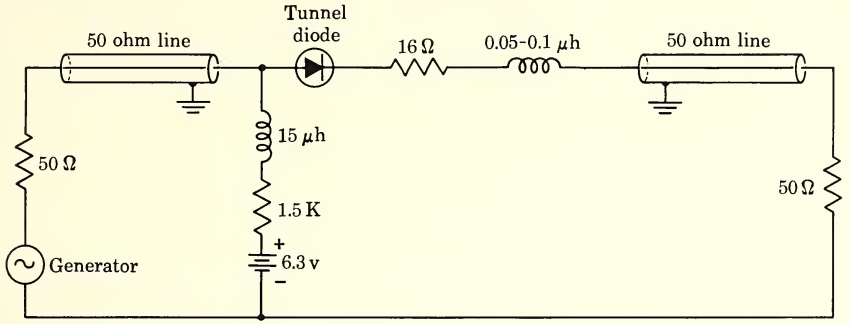


FIG. 11-9. Circuit of 100 Mc/sec amplifier.

CR_t/G, $\omega_R > \omega_X$ and the circuit becomes a nonlinear relaxation oscillator. This case is shown in the upper curve of Fig. 11-8. Sinusoidal oscillations occur when $\omega_R = \omega_X$, i.e., when $L_t = CR_t/G$.

Figure 11-9 shows the circuit of an amplifier having a bandwidth of 20 Mc/sec centered on 100 Mc/sec and a gain of 32 db. This shows the use of the tunnel diode as an amplifier at high frequencies.

In Fig. 11-10, a d-c load line has been drawn on the tunnel diode characteristic. The load line corresponds to the case where the total d-c load $R_1 < 1/G$, and there is only one intersection with the device characteristic. For the tunnel diode to operate reliably in this condition, the intersection point must be kept constant. Thus, E_1 must be accurately controlled.

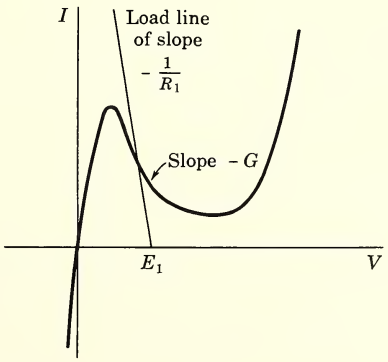


FIG. 11-10. Tunnel diode characteristic with load line where the load resistance is less than $-1/G$.

The load line drawn on Fig. 11-11(a) shows the case when $R_2 > 1/G$. This corresponds to the circuit of Fig. 11-11(b). There are three intersection points on the tunnel diode characteristic, A, B, and C. A and C are stable and B unstable. When a trigger voltage is placed in series with the supply, E_2, point A rises up the diode characteristic until it reaches the peak of the curve at a current I_p and a voltage V_p. The operating point then follows the

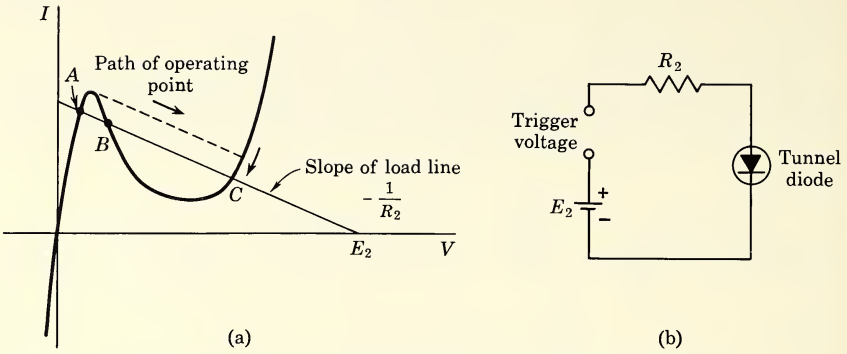


FIG. 11–11. Tunnel diode characteristic and circuit when used as a switch.

path shown by the dashed line. Thus the tunnel diode switches from A to C. A switching time of less than 10^{-9} sec can be achieved by a circuit of this type.

BIBLIOGRAPHY

Quantum mechanical tunneling

Bohm, David, *Quantum Theory*, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1951

Van Name, F. W. Jr., *Modern Physics*, 2d ed., Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1962

Tunnel diodes

Davidsohn, U.S., Y. C. Hwang, and G. B. Ober, "Designing with Tunnel Diodes," *Electronic Design*, **8**, 3 (February 3, 1960), 50–55; **8**, 4 (February 17, 1960), 66–71

Hall, R. N., "Tunnel Diodes," *I.R.E. Transactions on Electron Devices*, ED-7, 1 (January 1960), 1–9

Lesk, I. A. and J. J. Suran, "Tunnel Diode Operation and Application," *Electrical Engineering*, **79**, 4 (April 1960)

Todd, C. D., "Tunnel Diode Applications," *Electrical Engineering*, **80**, 4 (April 1961), 265–271

PROBLEMS

11–1 Derive the value of T as given in Eq. (11–11).

11–2 Calculate T for $U - E = 1$ ev where the particles are electrons and $a = 100$ A.

11–3 Draw the S-plane, Z-plane diagrams and a rough plot of R and X versus frequency for the three cases: $\omega_X < \omega_R$, $\omega_X = \omega_R$, and $\omega_X > \omega_R$.

11–4 Analyze the circuit of Fig. 11–9 showing that amplification will be obtained at a frequency of 100 Mc/sec and that stability is attained when L is less than 0.084 microhenries. For the tunnel diode: $R_s = 2$ ohms, $C = 5$ pf, $L_s = 5 \times 10^{-9}\,h$, $-G = -0.007$ mhos.

APPENDIX

Useful constants

Electronic charge e	1.602×10^{-19} coulombs
Electronic mass m	9.1085×10^{-31} kg
Planck's constant h	6.625×10^{-34} joule sec
Boltzmann's constant k	1.380×10^{-23} joule $°K^{-1}$
No. of atoms of germanium per cc	4.42×10^{22}
Permittivity of free space ϵ_0	$(36\pi \times 10^9)^{-1}$ farads m^{-1}
Relative permittivity of germanium	16
Relative permittivity of silicon	12

Index

A

Abrupt junctions, 72
a-c operation, 139–142
Acceptor state, 34
Active circuits, 144
Alloy-junction transistors, 129–130
Alpha-cutoff frequency, 194
Alternating-current gain, 127
Amplifiers, electronic devices as, 136–139
Amplifiers, single and multistage, 167–196
  biasing circuits, 178–182
  CE circuit, 167–196
  characteristic curves, 168–170
  class A, B, and C operation, 183
  frequency response of multistage, 194
  limitations of power amplifiers, 187–189
  load line, 170–172
  multistage, 191–193
  operating point temperature stability, 172–178
  power amplifiers, 182–186, 187–189
  thermal runaway, 189–191
Atomic binding, and valence, 17–18
Atomic number, 15
Atomic structure, wave-mechanical theory, 14–15
Atomic table, 15–17
Atomic table for 36 elements (table), 16
Avalanche breakdown, 86

B

β (beta), 128
Band theory, energy levels in crystals, 20–23
Bardeen, 113
Base charge control, concept of, 249
Base region, current flow across, 117–123
Base region, recombination in, 121
Base spreading resistance, 148, 203–204
Base width modulation, 120–121, 208–211
Beta-cutoff frequency, 194, 234
Bias, 137
Bias, in junctions, 79
Biasing, fixed-base, 176
Biasing circuits, 178–182
Body-centered cubic cell, 19
Bohr, Niels, hydrogen atom theory, 5–7, 10, 11, 14, 15, 31, 32, 34
Boltzmann factor, 48, 52, 99–100, 223, 264
Brattain, 113
"Built-in" field, 220–221, 223

C

CE circuit, 126–128, 167–196
Characteristics, static, 137
Charge control, 249–254
Charge depletion region, 73
Charges, mobile and immobile, 58–61
Collector-to-base cut-off current, 121–122
Collector current, 115–116
Collector multiplication, 187
Collector multiplication factor, 202–203
Common base circuit, 123–126
Common base connection, 236–238
Common emitter circuit (CE), 126–128, 167–196
Common emitter connection, frequency response in, 233–235
Conduction band, 22
Conductors, 25
Conductors, defined, 4
Contact potential, 72
Continuity equation, 89–111, 199–200
  abrupt *p-n* junctions with constant current, 94–99
  base region for direct currents, 200–202
  base region for small alternating signals, 204–208
  charge densities at edge of depletion regions, 99–101
  derivation, 90–94
  diffusion of charge, 91–94
  drift of charge in electric field, 91
  Einstein equation, 108–109
  electrons, equation for, 94
  electrons and positive holes, 91
  forward-biased abrupt *p-n* junctions, 104–107
  incremental resistance of forward-biased *p-n* junction, 108
  p-n junctions, 107
  positive holes, equation for, 93
  reverse-biased abrupt *p-n* junctions, 101–104
Covalent binding, 18
Crystal detectors, early, 112–113
Crystal unit cells, 19–20
Crystals, 18–20
Crystals, electronic mobility, 28–29
Current amplifier, transistor as, 139
Current gain, 124–128
Current negative feedback, 178
Cut-off current, 122

D

Davisson, 12
d-c conditions, 133–196
De Broglie, 12
Degeneracy, 264
Diamonds, 19, 20, 22, 23, 24
"Diffused-base transistor," term, 223
Diffusion, 91–94
Diffusion, current flow by, 117–119
Diffusion capacitance, 122–123, 213
Diffusion length, for electrons, 99
Diffusion length, for positive holes, 98
Diffusion techniques, in manufacturing, 130–131
Diffusion transistor, term, 220
Dirac, 12, 14
Distribution of power dissipation (table), 184
Donor state, 32–33
Doping, 35
Drift of charges, in electric fields, 91
Drift transistor, 220–223

E

Effective mass, concept of, 47
Effective width, of junctions, 73
Einstein equation, 108–109, 219, 222
Ekiss, J. A., 254
Electron energy levels, in crystals, 20–23
Electron volt, 10
Electronic mobility, in crystals, 28–29
Electrons:
  continuity equation for, 94
  diffusion length for, 99
  energy states, 15
  and positive holes, 26–28, 91
Elements, atomic table, 15–17
Emitter current, 115
Emitter efficiency, 119, 202
Emitter follower, 162
Emitter resistance, 213
Energy bands, density of states, 44–47
Energy density of states, defined, 47
Energy levels in crystals (band theory), 20–23
Energy state densities, and temperature, 47–51
Energy states, 15, 44–47
Epitaxial films, 226–227
Epitaxial mesa transistor, 227
Esaki, 260
Exclusion principle, Pauli, 17, 45, 48
Extrinsic semiconductors, 26

F

Fermi level, 48–58, 68–70, 72, 76, 100, 264–267
Fermi-Dirac statistics, 43, 70, 264
Fixed-base biasing, 176
Flicker noise, 223–224
Forbidden gap, 22, 25, 26, 27, 31, 44, 52, 53, 54, 56, 58, 128
Forward bias, 79
Frequency response, 229–259
  charge control, 249–254
  common base connection (simple model), 236–238
  common emitter connection, 233–235
  components of charge, 251–253
  hybrid-π circuits, 230–231, 233–236, 238–242, 244–247
  iterative common emitter stage, 235–236
  Miller effect, 231–232
  saturation, 242–244, 247–249
  switching, 244–247
  switching times, defined, 247
  switching times for saturation, 254–258

G

Gain-bandwidth product, 234
Gallium arsenide, 267
Generation, of electron-hole pairs, 35–40
Generation rate, 39
Germanium, 4, 18, 25, 31, 33, 128 (*also passim*)
Germanium and silicon, common donor and acceptor elements (table), 32
Germanium and silicon, properties of (table), 28
Germer, 12
Goudsmit, 14
Graded junctions, 72
Group 4 materials, 30
Group 5 substances, 30
Grown-junction transistors, 129

H

h, Planck's constant, 6, 14
h parameter equivalents of CE and CC circuits (table), 160
h parameter representation, low frequency, 154–166
  analysis of general equivalent circuit, 156–159
  common base circuit, 161–162
  common collector circuit, 162–163
  common emitter circuit, 163
  current gain, 158
  input resistance, 157
  output resistance, 158
  power gain, 158
  relationship of three sets of, 159–160
  small-signal, 155–156
  three configurations, comparison of, 161–163
  variation with I_E and V_{CE}, 163–165
  voltage gain, 157
h parameters, equivalent circuit using, 149–150
Hall effect, 61–64
Heat-sink, 191
Heisenberg, 12, 13–14, 45
Hybrid (*h*) parameters, defined, 149 (*see also h* parameter representation)
Hybrid-π circuit, 135, 143, 148–149, 200, 211–217, 229–259
Hydrogen atom:
  Bohr's theory of, 5–7
  energy of electron orbits, 7–9
  energy level representation, 10–11
  simple theory of, 4–5
  transitions between orbits, 9–10

I

Impedance terms, 141
Impurities, 18, 25–26, 29–35, 58
Incremental resistance, 108
Insulators, 24
Insulators, defined, 4
Intrinsic germanium, 60–61, 77
Intrinsic material, 29
Intrinsic semiconductors, 25, 34, 51–53
Intrinsic transistor, 203–204
Ionic binding, 18
Ionization energy, 10
Iterative common emitter stage, 235–236

J

Junction transistors, 112–133
Junctions:
  between metals, 70–72
  bias, 79
  breakdown region of *p-n* junctions, 86
  capacitance, 79–82
  effective width, 73
  flow of current across a *p-n* junction, 82–86
  incremental capacitance, 81
  nonlinear incremental capacitance, 81
  rectifier equation, 82–86
  with applied voltage, 79–82
  with no applied voltage, 72–78

L

Lattice constant, 20
Lifetimes, 38
Load line, 170–172
Low frequencies, 133–196
Low frequency measurements, 150–153

M

Majority carriers, 33, 34
Manufacture:
  alloy-junction transistors, 129–130
  diffusion techniques, 130–131
  grown-junction transistors, 129
  materials, 128–129
  point-contact transistors, 129
  zone refining, 128
Mass action, law of, 58–59
Matter, wave nature, 11–14
Maxwell-Boltzmann statistics, 52
Mesa diffused-base transistors, 130–131, 227
Metals, junctions between, 70–72
Metals, work function, 68–70
Miller effect, 231–232
Minority carrier lifetimes, 38
Minority carriers, 33, 34
Mobility of electrons and holes, 28–29, 36, 37
Modulation noise, 223
Momentum space diagram, 44, 46
Multistage amplifiers, 191–193 (*see also* Amplifiers)

N

npn transistors, 113–116
n- and *p*-type semiconductors, 29–35
n-type, term, 33
n-type germanium, 104
n-type material, 97–98
n-type semiconductors, 54–55
Noise, 223–225
Noise figure, defined, 224
Norton, 125, 146, 149, 150
Numbered and lettered *h* parameters (table), 155
Nyquist plot, for tunnel diode, 270

O

Ohmic contacts, 107
Operating point temperature stability, 172–178

P

π circuit, 148–149
Parameters, *z*, *y*, and *h*, 135, 141, 142–144
"Passivate," term, 226
Passive circuits, 144
Pauli exclusion principle, 17, 45, 48
Perfect crystals, 18
Performance quantities for CB, CE, and CC circuits (table), 147
Performance quantities for *h* parameter equivalent circuit (table), 158
Photon, 10
Physical characteristics, 199–228
  alpha, components of, 202–203
  base width modulation, 208–211
  continuity equation (*see* Continuity equation)
  drift transistor, 220–223
  epitaxial films, 226–227
  hole density in base region, 200–202
  hybrid-π representation, 211–217
  intrinsic transistor, 203–204
  noise, 223–225
  punch-through, 226
  small alternating signals, 204–208
  surface effects, 225–226
  transmission line analogy, 208–211
  variation of alpha with emitter current, 217–220
Planck, Max, quantum theory, 5–7, 14
p- and *n*-type semiconductors, 29–35
p-*n* junctions:
  breakdown region, 86
  conditions of operation, 96
  continuity equation (*see* Continuity equation)
  discussion of, 107
  flow of current across, 82–86
  as rectifiers, 35
pnp transistors, 116–117
  common base circuit, 123–126
  common emitter circuit, 126–128
  current flow across base region, 117–123
  hole density in base region, 200–202
Point-contact germanium diode, development of, 113
Point-contact transistors, development of, 113
Point-contact transistors, manufacture of, 129
Poisson equation, 74–76
Positive holes:
  continuity equation for, 93
  diffusion length for, 98
  effective mass of, 47
  and electrons, 26–28
Potential barriers, 78
Power amplifiers, 182–186
Properties of tunnel diodes (table), 267
p-type germanium, 104
p-type semiconductors, 55–56
Punch-through, 187–188, 226
Punch-through voltage, 226

Q

Quantum mechanical tunneling, 261–263
Quantum numbers, 14–15
Quantum theory, 5–7

R

Radiation, relation to matter, 11–14
Recombination, in base region, 121
Recombination, of electron-hole pairs, 35–40
Rectifier equation, 82–86
Relationship between parameters (table), 152
Resistance noise, 223
Reverse bias, 79
Rise time, defined, 247
Rutherford, 4

S

Saturation, 188, 242–244, 247–249
  current density, 83
  switching times for, 254–258
Schrödinger, 12–13, 15, 261, 263
Semiconductors:
  defined, 4
  electrical conduction in, 3–42
  electrons and holes, 43–66
Shockley, 113
Shot noise, 223
Silicon, 4, 18, 25, 128–129 (*also passim*)
Simmons, C. D., 254
Single crystals, 18
Solid-state devices, 4
Sommerfeld, 11, 68
Space charge region, 73
Speed-up capacitors, 257–258
"Spot" noise figure, 225
Stability factor equation, 177
Step junctions, 72
Storage time, 244
Surface effects, 225–226
Surface leakage, 187
Surface recombination velocity, 225–226
Switching, 244–247
Switching times, defined, 247
Switching times, for saturation, 254–258

T

T circuit, 144–147
Temperature:
  and energy state densities, 47–51
  limit placed by, 56–57
  stability, 172–178
  variation of Fermi level, 56–57
Thermal agitation, 223
Thermal equilibrium values, 35
Thermal resistance, 190
Thermal runaway, 172, 189–191
Thermionic emission, 70
Thévenin circuits, 146, 149, 150
Thomson, G. P., 12
Transmission line analogy, 208–211
Transport factor, 202
Traps (recombination centers), 38
Tunnel diode, 260–273
  as circuit element, 268–272
  energy bands, 263–265
  forward and reverse characteristics, 265–268
Tunnel diode (*con't*)
  limitations, 268
  Nyquist plot, 270
  oscillation frequencies, 268–269
  quantum mechanical tunneling, 261–263
  resistance cut-off angular frequency, 269
  self-resonant angular frequency, 269
  switching times, 268
Two port network, transistor as, 135–153
Typical h parameter values (table), 161

U

Uhlenbeck, 14
Uncertainty principle (Heisenberg), 13–14, 45
Unfilled energy states, 33
Unit momentum cells, 45

V

Vacuum level, 69
Valence, and atomic binding, 17–18
Valence, Fermi level, and work function for five metals (table), 70
Valence band, 22
Variation in I_{CBO}, h_{FE}, and I_B (table), 175
Voltage gain, 125, 126

W

Wave packet, 13
Wave (quantum) mechanics, 11–14
"White" noise, 223
Work function, of materials, 69
"Worst-case" design, 238

Y

y parameters, equivalent circuit using, 148–149

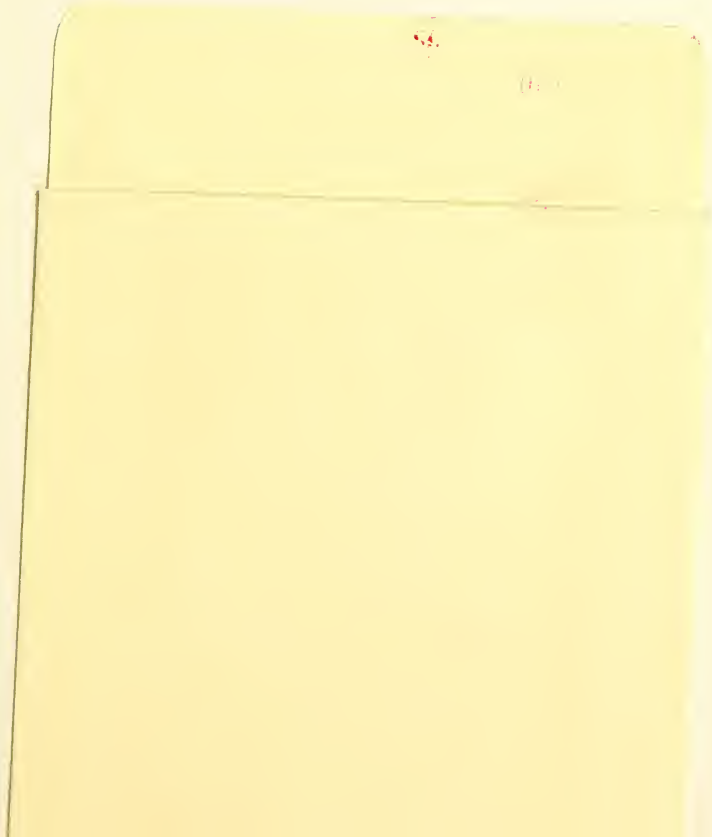
Z

z, y, and h parameters, 135, 141, 142–144
z parameters, equivalent circuit using, 144–147
Zener breakdown, 86
"Zener" devices, 86
Zone refining, 128

Date Due

Due	Returned	Due	Returned
DEC 6 '74	DEC 6 '74		
JAN 6 '75	DEC 13 '74		
MAY 23 '75	MAY 23 '75		
NOV 24 '76	NOV 30 '76		
APR 3 '78	MAR 15 '79		
OCT 18 1981	NOV 10 '81		
APR 30 '82	MAY 17 1982		
AUG 30 1982	AUG 31 1982		
SEP 21 1982	SEP 27 1982		
OCT 18 1982	OCT 21 1982		
JUN 08 1986	JUN 15 1986		
MAR 17 2000	AUG 28 2000		

Transistors engr
621.38151L469t C.2
3 1262 02205 5098

621.38151
L469t
C.2
ENGINEERING
& PHYSICS
LIBRARY

SCI-HSS